Multi-Task Averaging

Sergey Feldman, Maya R. Gupta, and Bela A. Frigyik
Department of Electrical Engineering
University of Washington
Seattle, WA 98103

Abstract

We present a multi-task learning approach to jointly estimate the means of multiple independent data sets. The proposed multi-task averaging (MTA) algorithm results in a convex combination of the single-task averages. We derive the optimal amount of regularization, and show that it can be effectively estimated. Simulations and real data experiments demonstrate that MTA outperforms both maximum likelihood and James-Stein estimators, and that our approach to estimating the amount of regularization rivals cross-validation in performance but is more computationally efficient.

1 Introduction

The motivating hypothesis behind multi-task learning (MTL) algorithms is that leveraging data from related tasks can yield superior performance over learning from each task independently. Early evidence for this hypothesis is Stein's work on the estimation of the means of T distributions (tasks) [1]. Stein showed that it is better (in a summed squared error sense) to estimate each of the means of T Gaussian random variables using data sampled from all of them, even if they are independent and have different means. That is, it is beneficial to consider samples from seemingly *unrelated* distributions in the estimation of the tth mean. This surprising result is often referred to as *Stein's paradox* [2].

Estimating means is perhaps the most common of all estimation tasks, and often multiple means need to be estimated. In this paper we consider a multi-task regularization approach to the problem of estimating multiple means that we call *multi-task averaging* (MTA). We show that MTA has provably nice theoretical properties, is effective in practice, and is computationally efficient. We define the MTA objective in Section 2, and review related work in Section 3. We present some key properties of MTA in Section 4 (proofs are omitted due to space constraints). In particular, we state the optimal amount of regularization to be used, and show that this optimal amount can be effectively estimated. Simulations in Section 5 verify the advantage of MTA over standard sample means and James-Stein estimation if the true means are close compared to the sample variance. In Section 6.1, two experiments estimating expected sales show that MTA can reduce real errors by over 30% compared to the sample mean. MTA can be used anywhere multiple averages are needed; we demonstrate this by applying it fruitfully to the averaging step of kernel density estimation in Section 6.1.

2 Multi-Task Averaging

Consider the T-task problem of estimating the means of T random variables that have finite mean and variance. Let $\{Y_{ti}\}_{i=1}^{N_t}$ be N_t independent and identically-distributed random samples for $t = 1, \ldots, T$. The MTA objective and many of the results in this paper generalize trivially to samples that are vectors rather than scalars, but for notational simplicity we restrict our focus to scalar samples $Y_{ti} \in \mathbb{R}$. Key notation is given in Table 1.

Table 1: Key Notation

T	number of tasks
N_t	number of samples for tth task
$Y_{ti} \in \mathbb{R}$	ith random sample from tth task
$\bar{Y}_t \in \mathbb{R}$	tth sample average $\frac{1}{N_t}\sum_i Y_{ti}$
$Y_t^* \in \mathbb{R}$	MTA estimate of tth mean
σ_t^2	variance of the tth task
Σ	diagonal covariance matrix of $\bar{Y}$ with $\Sigma_{tt} = \frac{\sigma_t^2}{N_t}$
$A \in \mathbb{R}^{T \times T}$	pairwise task similarity matrix
$L = D - A$	graph Laplacian of A, with diagonal D s.t. $D_{tt} = \sum_{r=1}^{T} A_{tr}$

In addition, assume that 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 without loss of generality (because the diagonal self-similarity terms are canceled in the objective below). The proposed MTA objective is

$$\{Y_t^*\}_{t=1}^T = \underset{\{\hat{Y}_t\}_{t=1}^T}{\arg\min} \ \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\frac{(Y_{ti}-\hat{Y}_t)^2}{\sigma_t^2} + \frac{\gamma}{T^2}\sum_{r=1}^{T}\sum_{s=1}^{T}A_{rs}(\hat{Y}_r-\hat{Y}_s)^2. \tag{1}$$

The first term minimizes the sum of the empirical losses, and the second term jointly regularizes the estimates by regularizing their pairwise differences. The regularization parameter γ balances the empirical risk and the multi-task regularizer. Note that if $\gamma = 0$, then (1) decomposes to T separate minimization problems, producing the sample averages $\bar{Y}_t$. The normalization of each error term in (1) by its task-specific variance σ_t^2 (which may be estimated) scales the T empirical loss terms relative to the variance of their distribution; this ensures that high-variance tasks do not disproportionately dominate the loss term.

A more general formulation of MTA is

$$\{Y_t^*\}_{t=1}^T = \underset{\{\hat{Y}_t\}_{t=1}^T}{\arg\min} \ \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N_t} L(Y_{ti},\hat{Y}_t) + \gamma J\left(\{\hat{Y}_t\}_{t=1}^T\right),$$

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 [3]. For the analysis and experiments in this paper, we restrict our focus to the tractable squared-error formulation given in (1).

The task similarity matrix A can be specified as side information (e.g. from a domain expert), or set in an optimal fashion. In Section 4 we derive two optimal choices of A for the $T = 2$ case: the A that minimizes expected squared error, and a minimax A. We use the $T = 2$ analysis to propose practical estimators of A for any number of tasks.

3 Related Work

MTA is an approach to the problem of estimating T means. We are not aware of other work in the multi-task literature that addresses this problem; most MTL methods are designed for regression, classification, or feature selection, e.g. [4, 5, 6]. The most closely related work is Stein estimation, an empirical Bayes strategy for estimating multiple means simultaneously [7, 8, 2, 9]. James and Stein [7] showed that the maximum likelihood estimate of the tth mean μ_t can be dominated by a shrinkage estimate given Gaussian assumptions. There have been a number of extensions to the original James-Stein estimator. We compare to the positive-part residual James-Stein estimator for multiple data points per task and independent unequal variances [8, 10], such that the estimated mean for the tth task is

$$\xi + \left(1 - \frac{T-3}{(\bar{Y}-\xi)^T\Sigma^{-1}(\bar{Y}-\xi)}\right)_+ (\bar{Y}_t - \xi), \tag{2}$$

where $(x)_+ = \max(0, x)$; Σ is a diagonal matrix of the estimated variances of each sample mean where $\Sigma_{tt} = \frac{\hat{\sigma}_t^2}{N_t}$ and the estimate is shrunk towards ξ, which is usually set to be the mean of the sample means (other choices are sometimes used) $\xi = \bar{\bar{Y}} = \frac{1}{T}\sum_t \bar{Y}_t$. Bock's formulation of (2) uses the *effective dimension* (defined as the ratio of the trace of Σ to the maximum eigenvalue of Σ) rather than the T in the numerator of (2) [8, 7, 10]. In preliminary practical experiments where Σ must be estimated from the data, we found that using the effective dimension significantly crippled the performance of the James-Stein estimator. We hypothesize that this is due to the high variance of the estimate of the maximum eigenvalue of Σ.

MTA can be interpreted as estimating means of T Gaussians with an intrinsic Gaussian Markov random field prior [11]. Unlike most work in graphical models, we do not assume any variables are conditionally independent, and generally have non-sparse inverse covariance.

A key issue for MTA and many other multi-task learning methods is how to estimate the similarity (or task relatedness) between tasks and/or samples if it is not provided. A common approach is to estimate the similarity matrix jointly with the task parameters [12, 13, 5, 14, 15]. For example, Zhang and Yeung [15] 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 a joint, alternating way. Applying such joint and alternating approaches to the MTA objective (1) leads to a degenerate solution with zero similarity. However, the simplicity of MTA enables us to specify the optimal task similarity matrix for $T = 2$ (see Sec. 4), which we generalize to obtain an estimator for the general multi-task case.

4 MTA Theory

For symmetric A with non-negative components[1], the MTA objective given in (1) is continuous, differentiable, and convex. It is straightforward to show that (1) has closed-form solution:

$$Y^* = \left(I + \frac{\gamma}{T}\Sigma L\right)^{-1} \bar{Y}, \tag{3}$$

where $\bar{Y}$ is the vector of sample averages with tth entry $\bar{Y}_t = \frac{1}{N_t}\sum_{i=1}^{N_t} Y_{ti}$, L is the graph Laplacian of A, and Σ is defined as before. With non-negative A and γ, the matrix inverse in (3) can be shown to always exist using the Gershgorin Circle Theorem [16].

Note that the (r, s)th entry of $\frac{\gamma}{T}\Sigma L$ goes to 0 as N_t approaches infinity, and since matrix inversion is a continuous operation, $\left(I + \frac{\gamma}{T}\Sigma L\right)^{-1} \to I$ in the norm. By the law of large numbers one can conclude that Y^* asymptotically approaches the true means.

4.1 Convexity of MTA Solution

From inspection of (3), it is clear that each of the elements of Y^* is a linear combination of the sample averages $\bar{Y}$. However, a stronger statement can be made:

Theorem: If $\gamma \geq 0$, $0 \leq A_{rs} < \infty$ for all r, s and $0 < \frac{\sigma_t^2}{N_t} < \infty$ for all t, then the MTA estimates $\{Y_t^*\}$ given in (3) are a convex combination of the task sample averages $\{\bar{Y}_t\}$.

Proof Sketch: The theorem requires showing that the matrix $W = \left(I + \frac{\gamma}{T}\Sigma L\right)^{-1}$ exists and is right-stochastic. Using the Gershgorin Circle Theorem [16], we can show that the real part of every eigenvalue of W^{-1} is positive. The matrix W^{-1} is a Z-matrix [17], and if the real part of each of the eigenvalues of a Z-matrix is positive, then its inverse has all non-negative entries (See Chapter 6, Theorem 2.3, G_{20}, and N_{38}, [17]). Finally, to prove that W has rows that sum to 1, first note that by definition the rows of the graph Laplacian L sum to zero. Thus $\left(I + \frac{\gamma}{T}\Sigma L\right)\mathbf{1} = \mathbf{1}$, and because we established invertibility, this implies the desired right-stochasticity: $\mathbf{1} = \left(I + \frac{\gamma}{T}\Sigma L\right)^{-1}\mathbf{1}$.

[1] If an asymmetric A is provided, using it with MTA is equivalent to using the symmetric $(A^T + A)/2$.

4.2 Optimal A for the Two Task Case

In this section we analyze the $T = 2$ task case, with N_1 and N_2 samples for tasks 1 and 2 respectively. Suppose $\{Y_{1i}\}$ are iid (independently and identically distributed) with finite mean μ_1 and finite variance σ_1^2, and $\{Y_{2i}\}$ are iid with finite mean $\mu_2 = \mu_1 + \Delta$ and finite variance σ_2^2. Let the task-relatedness matrix be $A = [0\ a; a\ 0]$, and without loss of generality, we fix $\gamma = 1$. Then the closed-form solution (3) can be simplified:

$$Y_1^* = \left(\frac{T + \frac{\sigma_2^2}{N_2} a}{T + \frac{\sigma_1^2}{N_1} a + \frac{\sigma_2^2}{N_2} a} \right) \bar{Y}_1 + \left(\frac{\frac{\sigma_1^2}{N_1} a}{T + \frac{\sigma_1^2}{N_1} a + \frac{\sigma_2^2}{N_2} a} \right) \bar{Y}_2. \tag{4}$$

It is straightforward to derive the mean squared error of Y_1^*:

$$\text{MSE}[Y_1^*] = \frac{\sigma_1^2}{N_1} \left(\frac{T^2 + 2T\frac{\sigma_2^2}{N_2} a + \frac{\sigma_1^2 \sigma_2^2}{N_1 N_2} a^2 + \frac{\sigma_2^4}{N_2^2} a^2}{(T + \frac{\sigma_1^2}{N_1} a + \frac{\sigma_2^2}{N_2} a)^2} \right) + \frac{\Delta^2 \frac{\sigma_1^4}{N_1^2} a^2}{(T + \frac{\sigma_1^2}{N_1} a + \frac{\sigma_2^2}{N_2} a)^2}. \tag{5}$$

Comparing to the MSE of the sample average, one obtains the following relationship:

$$\text{MSE}[Y_1^*] < \text{MSE}[\bar{Y}_1] \text{ if } \Delta^2 - \frac{\sigma_1^2}{N_1} - \frac{\sigma_2^2}{N_2} < \frac{4}{a}, \tag{6}$$

Thus the MTA estimate of the first mean has lower MSE if the squared mean-separation Δ^2 is small compared to the variances of the sample averages. Note that as a approaches 0 from above, the RHS of (6) approaches infinity, which means that a small amount of regularization can be helpful even when the difference between the task means Δ is large. Summarizing, if the two task means are close relative to each task's sample variance, MTA will help.

The risk is the sum of the mean squared errors: $\text{MSE}[Y_1^*]+\text{MSE}[Y_2^*]$, which is a convex, continuous, and differentiable function of a, and therefore the first derivative can be used to specify the optimal value a^*, when all the other variables are fixed. Minimizing $\text{MSE}[Y_1^*] + \text{MSE}[Y_2^*]$ w.r.t. a one obtains the following solution:

$$a^* = \frac{2}{\Delta^2}, \tag{7}$$

which is always non-negative.

Analysis of the second derivative shows that this minimizer always holds for the cases of interest (that is, for $N_1, N_2 \geq 1$). In the limit case, when the difference in the task means Δ goes to zero (while σ_t^2 stay constant), the optimal task-relatedness a^* goes to infinity, and the weights in (4) on $\bar{Y}_1$ and $\bar{Y}_2$ become $1/2$ each.

4.3 Estimating A from Data

Based on our analysis of the optimal A for the two-task case, we propose two methods to estimate A from data for arbitrary T. The first method is designed to minimize the approximate risk using a constant similarity matrix. The second method provides a minimax estimator. With both methods we can use the Sherman-Morrison formula to avoid taking the matrix inverse in (3), and the computation of Y^* is $O(T)$.

4.3.1 Constant MTA

Recalling that $E[\bar{Y}\bar{Y}^T] = \mu\mu^T + \Sigma$, the *risk* of estimator $\hat{Y} = W\bar{Y}$ of unknown parameter vector μ for the squared loss is the sum of the mean squared errors:

$$R(\mu, W\bar{Y}) = E[(W\bar{Y} - \mu)^T (W\bar{Y} - \mu)] = \textbf{tr}(W\Sigma W^T) + \mu^T (I - W)^T (I - W)\mu. \tag{8}$$

One approach to generalizing the results of Section 4.2 to arbitrary T is to try to find a symmetric, non-negative matrix A such that the (convex, differentiable) risk $R(\mu, W\bar{Y})$ is minimized for $W = \left(I + \frac{\gamma}{T}\Sigma L\right)^{-1}$ (recall L is the graph Laplacian of A). The problem with this approach is two-fold: (i) the solution is not analytically tractable for $T > 2$ and (ii) an arbitrary A has $T(T-1)$ degrees of freedom, which is considerably more than the number of means we are trying to estimate in the first place. To avoid these problems, we generalize the two-task results by constraining A to be a scaled constant matrix $A = a\mathbf{1}\mathbf{1}^T$, and find the optimal a^* that minimizes the risk in (8). In addition, w.l.o.g. we set γ to 1, and for analytic tractability we assume that all the tasks have the same variance, estimating Σ as $\frac{\mathbf{tr}(\Sigma)}{T}I$. Then it remains to solve:

$$a^* = \arg\min_a R\left(\mu, \left(I + \frac{1}{T}\frac{\mathbf{tr}(\Sigma)}{T}L(a\mathbf{1}\mathbf{1}^T)\right)^{-1}\bar{Y}\right),$$

which has the solution

$$a^* = \frac{2}{\frac{1}{T(T-1)}\sum_{r=1}^{T}\sum_{s=1}^{T}(\mu_r - \mu_s)^2},$$

which reduces to the optimal two task MTA solution (7) when $T = 2$. In practice, one of course does not have $\{\mu_r\}$ as these are precisely the values one is trying to estimate. So, to estimate a^* we use the sample means $\{\bar{y}_r\}$: $\hat{a}^* = \frac{2}{\frac{1}{T(T-1)}\sum_{r=1}^{T}\sum_{s=1}^{T}(\bar{y}_r - \bar{y}_s)^2}$. Using this estimated optimal *constant* similarity and an estimated covariance matrix $\hat{\Sigma}$ produces what we refer to as the *constant MTA* estimate

$$Y^* = \left(I + \frac{1}{T}\hat{\Sigma}L(\hat{a}^*\mathbf{1}\mathbf{1}^T)\right)^{-1}\bar{Y}. \tag{9}$$

Note that we made the assumption that the entries of Σ were the same in order to be able to derive the constant similarity a^*, but we do not need nor suggest that assumption on the $\hat{\Sigma}$ used with $\hat{a}^*$ in (9).

4.4 Minimax MTA

Bock's James-Stein estimator is *minimax* in that it minimizes the worst-case loss, not necessarily the expected loss [10]. This leads to a more conservative use of regularization. In this section, we derive a minimax version of MTA, that prescribes less regularization than the constant MTA. Formally, an estimator Y^M of μ is called minimax if it minimizes the maximum risk:

$$\inf_{\hat{Y}}\sup_{\mu} R(\mu, \hat{Y}) = \sup_{\mu} R(\mu, Y^M).$$

First, we will specify minimax MTA for the $T = 2$ case. To find a minimax estimator Y^M it is sufficient to show that *(i)* Y^M is a Bayes estimator w.r.t. the least favorable prior (LFP) and *(ii)* it has constant risk [10]. To find a LFP, we first need to specify a constraint set for μ_t; we use an interval: $\mu_t \in [b_l, b_u]$, for all t, where $b_l \in \mathbb{R}$ and $b_u \in \mathbb{R}$. With this constraint set the minimax estimator is:

$$Y^M = \left(I + \frac{2}{T(b_u - b_l)^2}\Sigma L(\mathbf{1}\mathbf{1}^T)\right)^{-1}\bar{Y}, \tag{10}$$

which reduces to (7) when $T = 2$. This minimax analysis is only valid for the case when $T = 2$, but we found that good practical results for larger T using (10) with the data-dependent interval $\hat{b}_l = \min_t \bar{y}_t$ and $\hat{b}_u = \max_t \bar{y}_t$.

5 Simulations

We first illustrate the performance of the proposed MTA using Gaussian and uniform simulations so that comparisons to ground truth can be made. Simulation parameters are given in the table in Figure 1, and were set so that the variances of the distribution of the true means were the same in both types of simulations. Simulation results are reported in Figure 1 for different values of σ_μ^2, which determines the variance of the distribution over the means.

We compared constant MTA and minimax MTA to single-task sample averages and to the James-Stein estimator given in (2). We also compared to a randomized 5-fold 50/50 cross-validated (CV) version of constant MTA, and minimax MTA, and the James-Stein estimator (which is simply a convex regularization towards the average of the sample means: $\lambda\bar{y}_t + (1-\lambda)\bar{\bar{y}}.$). For the cross-validated versions, we randomly subsampled $N_t/2$ samples and chose the value of γ for constant/minimax

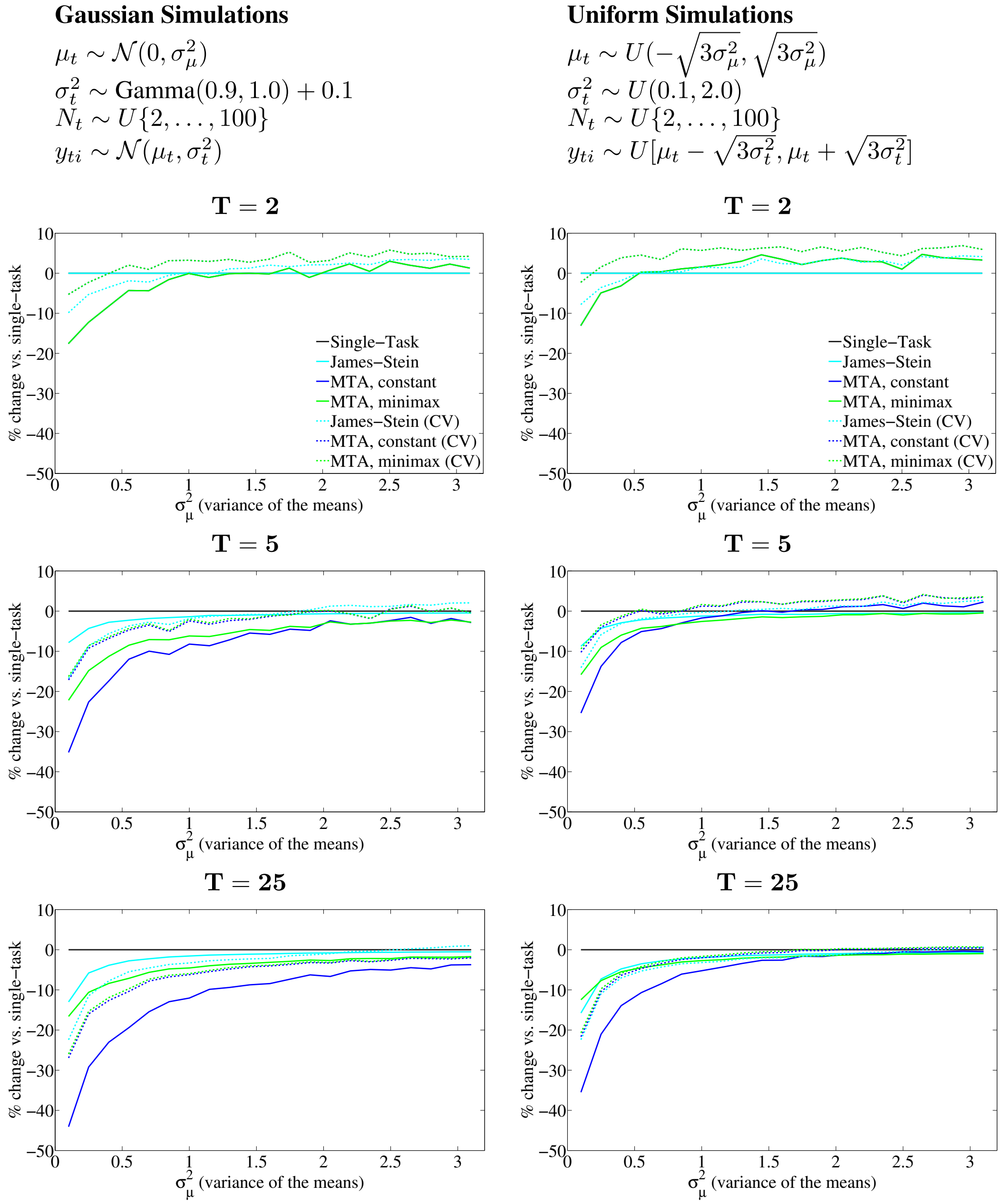

Figure 1: Average (over 10000 random draws) percent change in risk vs. single-task. Lower is better.

MTA or λ for James-Stein that resulted in the lowest average left-out risk compared to the sample mean estimated with *all* N_t samples. In the optimal versions of constant/minimax MTA, γ was set to 1, as this was the case during derivation.

We used the following parameters for CV: $\gamma \in \{2^{-5}, 2^{-4}, \ldots, 2^5\}$ for the MTA estimators and a comparable set of λ spanning $(0, 1)$ by the transformation $\lambda = \frac{\gamma}{\gamma+1}$. Even when cross-validating, an advantage of using the proposed constant MTA or minimax MTA is that these estimators provide a data-adaptive scale for γ, where $\gamma = 1$ sets the regularization parameter to be $\frac{a^*}{T}$ or $\frac{1}{T(b_u - b_l)^2}$, respectively.

Some observations from Figure 1: further to the right on the x-axis, the means are more likely to be further apart, and multi-task approaches help less on average. For $T = 2$, the James-Stein estimator reduces to the single-task estimator, and is of no help. The MTA estimators provide a gain while

$\sigma_\mu^2 < 1$ but deteriorates quickly thereafter. For $T = 5$, constant MTA dominates in the Gaussian case, but in the uniform case does worse than single-task when the means are far apart. Note that for all $T > 2$ minimax MTA almost always outperforms James-Stein and always outperforms single-task, which is to be expected as it was designed conservatively. For $T = 25$, we see the trend that all estimators benefit from an increase in the number of tasks.

For constant MTA, cross-validation is always worse than the estimated optimal regularization. Since both constant MTA and minimax MTA use a similarity matrix of all ones scaled by a constant, cross-validating over a set of possible γ may result in nearly identical performance, and this can be seen in the Figure (i.e. the green and blue dotted lines are superimposed). To conclude, when the tasks are close to each other compared to their variances, constant MTA is the best estimator to use by a wide margin. When the tasks are farther apart, minimax MTA will provide a win over both James-Stein and maximum likelihood.

6 Applications

We present two applications with real data. The first application parallels the simulations, estimating expected values of sales of related products. The second application uses MTA for multi-task kernel density estimation, highlighting the applicability of MTA to any algorithm that uses sample averages.

6.1 Application: Estimating Product Sales

We consider two multi-task problems using sales data over a certain time period supplied by Artifact Puzzles, a company that sells jigsaw puzzles online. For both problems, we model the given samples as being drawn iid from each task.

The first problem estimates the impact of a particular *puzzle* on repeat business: "Estimate how much a random customer will spend on an order on average, if on their last order they purchased the tth puzzle, for each of $T = 77$ puzzles." The samples were the amounts different customers had spent on orders after buying each of the t puzzles, and ranged from 480 down to 0 for customers that had not re-ordered. The number of samples for each puzzle ranged from $N_t = 8$ to $N_t = 348$.

The second problem estimates the expected order size of a particular *customer*: "Estimate how much the tth customer will spend on a order on average, for each of the $T = 477$ customers that ordered at least twice during the data timeframe." The samples were the order amounts for each of the T customers. Order amounts varied from 15 to 480. The number of samples for each customer ranged from $N_t = 2$ to $N_t = 17$.

There is no ground truth. As a metric to compare the estimates, we treat each task's sample average computed from all of the samples as the ground truth, and compare to estimates computed from a uniformly randomly chosen 50% of the samples. Results in Table 2 are averaged over 1000 random draws of the 50% used for estimation. We used 5-fold cross-validation with the same parameter choices as in the simulations section.

Table 2: Percent change in average risk (for puzzle and buyer data, lower is better), and mean reciprocal rank (for terrorist data, higher is better).

Estimator	Puzzles	Customers	Suicide Bombings
	$T = 77$	$T = 477$	$T = 7$
Pooled Across Tasks	181.67%	109.21%	0.13
James-Stein	-6.87%	-14.04%	0.15
James-Stein (CV)	-21.18%	-31.01%	0.15
Constant MTA	-17.48%	**-32.29%**	0.19
Constant MTA (CV)	**-21.65%**	-30.89%	0.19
Minimax MTA	-8.41%	-2.96%	0.19
Minimax MTA (CV)	-19.83 %	-25.04%	0.19
Expert MTA	-	-	0.19
Expert MTA (CV)	-	-	0.19

6.2 Density Estimation for Terrorism Risk Assessment

MTA can be used whenever multiple averages are taken. In this section we present multi-task kernel density estimation, as an application of MTA. Recall that for standard single-task kernel density estimation (KDE) [18], a set of random samples $x_i \in \mathbb{R}^d, i \in \{1, \ldots, N\}$ are assumed to be iid from an unknown distribution p_X, and the problem is to estimate the density for a query sample, $z \in \mathbb{R}^d$. Given a kernel function $K(x_i, x_j)$, the un-normalized single-task KDE estimate is $\hat{p}(z) = \frac{1}{N}\sum_{i=1}^{N} K(x_i, z)$, which is just a sample average.

When multiple kernel densities $\{p_t(z)\}_{t=1}^{T}$ are estimated for the same domain, we replace the multiple sample averages with MTA estimates, which we refer to as multi-task kernel density estimation (MT-KDE).

We compared 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[2] 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, [19] also used a Gaussian kernel density estimate to assess risk from past terrorism events.

The goal in this application is to 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 17 suicide bombings. All the events are attributed to one of seven terrorist groups. The density estimates for these seven groups are expected to be related, and are treated as $T = 7$ tasks.

The kernel K was taken to be a Gaussian kernel with identity covariance. In addition to constant A and minimax A, we also obtained a side-information A from terrorism expert Mohammed M. Hafez of the Naval Postgraduate School; he assessed the similarity between the seven groups during the Second Intifada (the time period of the data), providing similarities between 0 and 1.

We used leave-one-out cross validation to assess KDE and MT-KDE for this problem, as follows. 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 mean reciprocal rank (MRR) metric is reported in Table 2. Ideally, the MRR of the left-out events would be as close to 1 as possible, and indicating that the location of the left-out event is at high-risk. The results show that the MRR for MT-KDE are lower or not worse than those for KDE for both problems; there are, however, too few samples to verify statistical significance of these results.

7 Summary

Though perhaps unintuitive, we showed that both in theory and in practice, estimating multiple *unrelated* means using an MTL approach can improve the overall risk, even more so than James-Stein estimation. Averaging is common, and MTA has potentially broad applicability as a subcomponent in many algorithms, such as k-means clustering, kernel density estimation, or non-local means denoising.

Acknowledgments

We thank Peter Sadowski, Mohammed Hafez, Carol Chang, Brian Sandberg and Ruth Wilis for help with preliminary experiments and access to the terorrist dataset.

[2] Primary sources included the NRL Israel Suicide Terrorism Database (ISD) cross referenced with 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 Institute for the Study of Violent Groups, the University of Maryland Global Terrorism Database, and the National Counter Terrorism Center Worldwide Incident Tracking System.

References

[1] C. Stein, “Inadmissibility of the usual estimator for the mean of a multivariate distribution,” *Proc. Third Berkeley Symposium on Mathematical Statistics and Probability*, pp. 197–206, 1956.

[2] B. Efron and C. N. Morris, “Stein’s paradox in statistics,” *Scientific American*, vol. 236, no. 5, pp. 119–127, 1977.

[3] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, “Clustering with Bregman divergences,” *Journal Machine Learning Research*, vol. 6, pp. 1705–1749, December 2005.

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

[5] 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.

[6] A. Argyriou, T. Evgeniou, and M. Pontil, “Convex multi-task feature learning,” *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[7] W. James and C. Stein, “Estimation with quadratic loss,” *Proc. Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 361–379, 1961.

[8] M. E. Bock, “Minimax estimators of the mean of a multivariate normal distribution,” *The Annals of Statistics*, vol. 3, no. 1, 1975.

[9] G. Casella, “An introduction to empirical Bayes data analysis,” *The American Statistician*, pp. 83–87, 1985.

[10] E. L. Lehmann and G. Casella, *Theory of Point Estimation*. New York: Springer, 1998.

[11] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*, ser. Monographs on Statistics and Applied Probability. London: Chapman & Hall, 2005, vol. 104.

[12] 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.

[13] Y. Xue, X. Liao, L. Carin, and B. Krishnapuram, “Multi-task learning for classification with Dirichlet process priors,” *Journal Machine Learning Research*, vol. 8, pp. 35–63, 2007.

[14] L. Jacob, F. Bach, and J.-P. Vert, “Clustered multi-task learning: A convex formulation,” in *Advances in Neural Information Processing Systems (NIPS)*, 2008, pp. 745–752.

[15] 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.

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

[17] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, 1979.

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

[19] D. Brown, J. Dalton, and H. Hoyle, “Spatial forecast methods for terrorist events in urban environments,” *Lecture Notes in Computer Science*, vol. 3073, pp. 426–435, 2004.