Statistica Sinica Preprint No: SS-2026-0338	
Title	Multitask Smoothed Support Vector Machines for High-dimensional Heterogeneous Data
Manuscript ID	SS-2026-0338
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202026.0338
Complete List of Authors	Ling Peng, Yue Wang, Xiaohui Liu and Heng Lian
Corresponding Authors	Yue Wang
E-mails	yuewangcityu@163.com
Notice: Accepted author version.	

Statistica Sinica: Newly accepted Paper

Multitask Smoothed Support Vector Machines for High-Dimensional Heterogeneous Data

Ling Peng[1], Yue Wang[2], Xiaohui Liu[1] and Heng Lian[3,4]

[1] *Jiangxi University of Finance and Economics*

[2] *University of Science and Technology of China*

[3] *City University of Hong Kong Shenzhen Research Institute*

[4] *City University of Hong Kong*

Abstract: We study high-dimensional binary classification across multiple related but heterogeneous tasks. For each task, the proposed method models the task-specific coefficient vector of a linear support vector machine (SVM) as the sum of a sparse common component and a row-sparse task-specific deviation, thereby separating features shared across tasks from those that capture cross-task heterogeneity. To handle the nonsmooth hinge loss, we employ a convolution-smoothing technique and estimate the common and deviation components through a composite regularizer that combines an ℓ_1 penalty with a row-wise $\ell_{2,1}$ penalty. Theoretically, we establish nonasymptotic estimation error bounds under high-dimensional settings, with the rate explicitly decomposing the stochastic error from the smoothing bias. We further prove variable selection consistency for both the shared active set and the heterogeneous active row set under irrepresentability and beta-min conditions. For computation, we develop a projected proximal-gradient algorithm tailored to the constrained composite optimization problem. Simulation studies and an application to PISA data illustrate the finite-sample performance of the proposed method relative to pooled and task-specific alternatives.

Key words and phrases: Convolution-smoothing, heterogeneous data, high-dimensional classification, multitask learning, support vector machine, variable selection.

Statistica Sinica: Newly accepted Paper

1. Introduction

Support vector machines (SVMs) are among the most widely used margin-based methods for binary classification. Since the seminal work of Cortes and Vapnik (1995), SVMs have been studied extensively from both algorithmic and statistical perspectives. The empirical hinge loss serves as a convex surrogate for the misclassification error, while margin maximization provides a clear geometric interpretation of the resulting classifier; see also Vapnik (2000); Cristianini and Shawe-Taylor (2000); Schölkopf and Smola (2002); Steinwart and Christmann (2008); Wang and Lian (2024). Classical SVM theory has focused primarily on prediction consistency and the risk properties of kernel classifiers. For linear SVMs, Koo et al. (2008) established a Bahadur-type representation in fixed dimensions, while Park et al. (2012) studied the asymptotic properties of SCAD-penalized SVMs in growing dimensions, and Peng et al. (2016) derived error bounds for ℓ_1-regularized SVM coefficients in ultra-high dimensions. Extensions to structured sparsity include the fused-lasso SVM of Tibshirani et al. (2005), the doubly regularized SVM of Wang et al. (2006), and distributed SVM learning methods such as that of Lian and Fan (2018); Wang et al. (2026). Although considerable progress has been made, these methods remain largely confined to either a single classification task or multiple data sources from the same underlying task.

Many contemporary applications, however, frequently go beyond this setting, where multiple classification tasks are related yet heterogeneous. For example, classification rules may share a common predictive structure across populations, institutions, or geographical regions, while still exhibiting systematic task-specific differences. Analyzing each task separately fails to exploit their common structure, whereas pooling all tasks imposes complete homogeneity and may lead to substantial bias. Multitask learning seeks a principled compromise between

these two extremes by borrowing strength across related tasks while preserving task-specific heterogeneity. There is a rich literature on algorithmic and regularization frameworks for multitask learning. Evgeniou and Pontil (2004) proposed a regularized multitask learning framework, and Evgeniou et al. (2005) subsequently developed multitask kernel methods that extend standard single-task procedures, including SVMs, through kernels and regularizers that encode task relatedness. Other influential approaches include learning predictive structures shared across tasks (Ando and Zhang, 2005), shared sparse representations (Argyriou et al., 2008), clustered task structures (Jacob et al., 2008), and low-rank or group-sparse coefficient matrices (Chen et al., 2010, 2011). For a comprehensive review, see Zhang and Yang (2022). These works demonstrate that margin-based and regularization-based multitask procedures can improve prediction when tasks are related and each task has limited data. Nevertheless, the statistical foundations of multitask SVMs in high-dimensional settings remain largely underexplored.

This paper addresses exactly this gap. We consider a different and practically relevant statistical regime, where linear decision rules are desirable for their interpretability, the ambient dimension may greatly exceed the task-specific sample sizes, and the underlying classification rules are neither identical nor completely unrelated. Specifically, we model each task-specific coefficient vector as a sum of a sparse common component and a row-sparse task-specific deviation. The common component draws on sparse high-dimensional estimation theory for ℓ_1 regularization (Bickel et al., 2009; Negahban et al., 2012; Wainwright, 2019), while the row-sparse deviation component is related to group and multiresponse regularization (Yuan and Lin, 2006; Meier et al., 2008; Obozinski et al., 2011). This decomposition allows us to distinguish features that are predictive across tasks from features that explain

Statistica Sinica: Newly accepted Paper

task heterogeneity.

Our work is also related to integrative analysis of heterogeneous high-dimensional data. A large literature has studied regularized estimation when several studies, sites, or subpopulations share partial structure but are not fully homogeneous. Examples include transfer learning for high-dimensional regression (Li et al., 2022), DataSHIELD-type analysis under individual-data-sharing restrictions (Wolfson et al., 2010), and heterogeneity-aware distributed inference (Duan et al., 2022). In particular, Cai et al. (2022) proposed SHIR, an individual-data-protected integrative regression method for heterogeneous studies, and developed rigorous estimation and variable-selection theory using summary statistics. Their framework, however, involves smooth regression losses and therefore does not directly apply to SVM classification. Bridging this gap requires addressing several challenges that are specific to high-dimensional multitask SVMs. **The nondifferentiability of the empirical hinge loss complicates the gradient- and Hessian-based arguments used in our analysis. We employ convolution smoothing to facilitate computation and the proof of variable selection consistency.** We therefore employ the convolution smoothing technique used in Tan et al. (2022); Wang et al. (2023); Rybak et al. (2025); Wang et al. (2026) to construct a differentiable approximation to the hinge loss. This step introduces a new difficulty: the curvature of the smoothed population risk depends on the distribution of the covariates near the margin boundary, while the smoothing procedure generates a bandwidth-dependent approximation bias. A valid high-dimensional analysis must therefore simultaneously control the empirical stochastic error, the smoothing bias, and the interaction between the sparse common component and the row-sparse deviation matrix. Although the resulting bias–variance tradeoff is related in spirit to smoothing methods for nonsmooth

estimating problems (Horowitz, 1992; Fernandes et al., 2021; Tan et al., 2022), the composite shared-and-heterogeneous structure considered here calls for new blockwise and active-set theoretical tools. This combination of a nonsmooth margin loss, high-dimensional sparsity, and multitask heterogeneity distinguishes our theoretical problem from both existing multitask SVM methods and high-dimensional integrative regression.

In addition, existing multitask SVM and multitask kernel methods are primarily designed as general learning algorithms or general regularization frameworks. They typically do not provide high-dimensional coefficient error bounds or support recovery theory in the context of heterogeneous sparse linear SVMs. The available high-dimensional multitask theory is mostly developed for squared-loss or generalized regression settings, including multitask group lasso and related matrix-regularized estimators (Obozinski et al., 2011; Negahban et al., 2012). In contrast, our goal is to provide a statistical theory for a composite-regularized multitask smoothed SVM, covering both estimation error bounds and variable selection consistency for this nonsmooth margin-based problem.

The main contributions of this paper are summarized as follows. First, we propose a multitask smoothed SVM estimator equipped with a composite penalty that combines an ℓ_1 penalty on the common component and an $\ell_{2,1}$ penalty on task-specific deviations. The resulting formulation explicitly captures shared and heterogeneous predictive structures while remaining computationally tractable. Second, we develop nonasymptotic error bounds for the estimator under high-dimensional scaling. The bounds explicitly quantify the contributions of stochastic estimation error and smoothing bias. Third, we establish variable selection consistency for recovering both the shared active set and the heterogeneous active row set under irrepresentability and beta-min conditions. Finally, we provide an efficient

Statistica Sinica: Newly accepted Paper

optimization algorithm and conduct extensive numerical experiments to illustrate the practical behavior of our method. The results demonstrate its advantages over both single-task and pooled baselines.

1.1 Organization and Notations

The rest of this article is organized as follows. Section 2 introduces the multitask smoothed SVM framework, including the common-deviation decomposition and the associated composite regularization. Section 3 presents the main theoretical results on estimation error and variable selection. Section 4 reports simulation studies and Section 5 presents the real data application. Section 6 concludes the article with a discussion. All technical proofs are provided in the Supplementary Material.

Notations. For any positive integer p, let $[p] = \{1, \ldots, p\}$. For two variables, $a \vee b$ denotes $\max\{a, b\}$. For a vector $\mathbf{v} = (v_1, \ldots, v_n)^\top$, define its ℓ_d norm by $\|\mathbf{v}\|_d = (\sum_{i=1}^n |v_i|^d)^{1/d}$ for $d \geq 1$, and write $\mathbf{v}_{-1} = (v_2, \ldots, v_n)^\top$. We abbreviate $\|\mathbf{v}\|_2$ as $\|\mathbf{v}\|$. For $S \subseteq [n]$, $\mathbf{v}_S$ is the subvector of $\mathbf{v}$ with components indexed by S. For a matrix $\mathbf{A}$ with entries a_{ij}, $\|\mathbf{A}\|$ denotes its Frobenius norm, $\|\mathbf{A}\|_{op}$ its operator norm, and $\|\mathbf{A}\|_\infty = \max_i \sum_j |a_{ij}|$ the matrix norm induced by the vector ℓ_∞ norm. For $\mathbf{A} \in \mathbb{R}^{p \times M}$, let $\|\mathbf{A}\|_{2,1} = \sum_{j=1}^p \|\mathbf{A}_{j\cdot}\|_2$. For a symmetric matrix $\mathbf{A}$, $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ denote its smallest and largest eigenvalues. For positive sequences a_n and b_n, write $a_n \lesssim b_n$ if $a_n \leq C b_n$ for a constant $C > 0$, $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $b_n \lesssim a_n$, and $a_n = o(b_n)$ if $a_n / b_n \to 0$. For a set A, $|A|$ denotes its cardinality, and $I(\cdot)$ denotes the indicator function. Superscripts of the form $^{(m)}$ index tasks. Throughout the paper, $C, C_1, C_2, \ldots$, denote generic positive constants whose values may change even on the same line.

2. Multitask Smoothed SVM with Composite Regularization

2.1 Problem Formulation

We consider a multitask binary classification problem with M related tasks. For each task $m \in [M]$, we observe independent and identically distributed (i.i.d.) data $\{(\mathbf{x}_i^{(m)}, y_i^{(m)})\}_{i=1}^{n_m}$, where $\mathbf{x}_i^{(m)} \in \mathbb{R}^p$ and $y_i^{(m)} \in \{-1, 1\}$. Denote the total sample size by $N = \sum_{m=1}^{M} n_m$. To capture both shared and task-specific effects, we decompose the task-specific slope vector as

$$\boldsymbol{\beta}^{(m)} = \boldsymbol{\mu} + \boldsymbol{\alpha}^{(m)},$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is a common sparse component shared across all tasks and $\boldsymbol{\alpha}^{(m)} \in \mathbb{R}^p$ represents task-dependent deviations. We also allow task-specific intercepts $b^{(m)} \in \mathbb{R}$ and denote $\mathbf{b} = (b^{(1)}, \ldots, b^{(M)})^\top$. To ensure the identifiability of the decomposition, we impose the centering constraint $\mathbf{1}^\top \boldsymbol{\alpha}_j = 0$ for each j, where $\boldsymbol{\alpha}_j = (\alpha_j^{(1)}, \ldots, \alpha_j^{(M)})^\top$ collects the j-th coordinates of $\{\boldsymbol{\alpha}^{(m)}\}_{m=1}^{M}$. This centering convention makes $\boldsymbol{\mu}$ the across-task average effect and makes $\boldsymbol{\alpha}^{(m)}$ a genuine deviation from that average. Without such a normalization, the same collection of task-specific coefficients $\{\boldsymbol{\beta}^{(m)}\}_{m=1}^{M}$ could be represented by infinitely many pairs $(\boldsymbol{\mu}, \mathbf{A})$, which would make support recovery for the shared and heterogeneous components ill-defined. Let $\mathbf{A} := (\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(M)}) \in \mathbb{R}^{p \times M}$ denote the task-specific deviation matrix, whose m-th column corresponds to the deviation vector of task m. Let $\|\mathbf{A}\|_{2,1} = \sum_{j=1}^{p} \sqrt{\sum_{m=1}^{M} (\alpha_j^{(m)})^2}$ be the feature-wise group norm across tasks. The estimator is obtained by minimizing the

Statistica Sinica: Newly accepted Paper

multitask empirical hinge loss with a composite regularizer:

$$
\begin{aligned}
&(\widetilde{\mathbf{b}}, \widetilde{\boldsymbol{\mu}}, \widetilde{\mathbf{A}}) \\
&:= \underset{\substack{\mathbf{b},\boldsymbol{\mu},\mathbf{A} \\ \mathbf{1}^\top \boldsymbol{\alpha}_j = 0,\ j \in [p]}}{\operatorname{argmin}} \left\{ \frac{1}{N} \sum_{m=1}^{M} \sum_{i=1}^{n_m} \phi\left(y_i^{(m)} \left((\mathbf{x}_i^{(m)})^\top (\boldsymbol{\mu} + \boldsymbol{\alpha}^{(m)}) + b^{(m)} \right) \right) \right. \\
&\qquad \left. + \lambda \left(\|\boldsymbol{\mu}\|_1 + \lambda_g \|\mathbf{A}\|_{2,1} \right) \right\}, \qquad (2.1)
\end{aligned}
$$

where $\phi(z) = (1-z)_+$ is the hinge loss and $\lambda, \lambda_g > 0$ are tuning parameters. Here, we impose a composite regularizer consisting of an ℓ_1 and an $\ell_{2,1}$ penalty. The $\ell_{2,1}$ penalty encourages feature-wise group sparsity in the task-specific deviations, so that heterogeneity across tasks is concentrated on a small subset of features, whereas the ℓ_1 penalty promotes sparsity in the shared component. The two penalties play complementary roles. If a feature has a common effect across many tasks, it is more economically represented through $\boldsymbol{\mu}$ and selected by the ℓ_1 penalty. If its effect varies across tasks, the corresponding row of $\mathbf{A}$ is selected as a group, so the method can identify which covariates are responsible for heterogeneity rather than treating every task-specific coefficient separately.

The unregularized population counterpart of (2.1) is defined as

$$
\begin{aligned}
&(\mathbf{b}^*, \boldsymbol{\mu}^*, \mathbf{A}^*) \\
&\in \underset{\substack{\mathbf{b},\boldsymbol{\mu},\mathbf{A} \\ \mathbf{1}^\top \boldsymbol{\alpha}_j = 0,\ j \in [p]}}{\operatorname{argmin}} \left\{ \frac{1}{N} \sum_{m=1}^{M} n_m \mathbb{E}^{(m)} \left[\phi\left(y^{(m)} \left((\mathbf{x}^{(m)})^\top (\boldsymbol{\mu} + \boldsymbol{\alpha}^{(m)}) + b^{(m)} \right) \right) \right] \right\}, \qquad (2.2)
\end{aligned}
$$

where $\mathbb{E}^{(m)}$ denotes the expectation with respect to the task-m population distribution. We regard $(\mathbf{b}^*, \boldsymbol{\mu}^*, \mathbf{A}^*)$ (assumed to uniquely exist) as the target parameter. For each task

Statistica Sinica: Newly accepted Paper

$m = 1, \ldots, M$, define

$$\widetilde{\mathbf{x}}^{(m)} = (1, (\mathbf{x}^{(m)})^\top)^\top \in \mathbb{R}^{p+1}, \qquad \boldsymbol{\theta}^{(m)} := \left(b^{(m)}, (\boldsymbol{\mu} + \boldsymbol{\alpha}^{(m)})^\top\right)^\top \in \mathbb{R}^{p+1},$$

and define the task-specific population score with respect to the augmented parameter $\boldsymbol{\theta}^{(m)}$ by

$$\mathbf{s}^{(m)}(\boldsymbol{\theta}^{(m)}) := -\mathbb{E}^{(m)}\left[I\left\{1 - y^{(m)}(\widetilde{\mathbf{x}}^{(m)})^\top \boldsymbol{\theta}^{(m)} \geq 0\right\} y^{(m)} \widetilde{\mathbf{x}}^{(m)}\right].$$

Similarly, the corresponding Hessian matrix is

$$\mathbf{H}^{(m)}(\boldsymbol{\theta}^{(m)}) := \mathbb{E}^{(m)}\left[\delta\left(1 - y^{(m)}(\widetilde{\mathbf{x}}^{(m)})^\top \boldsymbol{\theta}^{(m)}\right) \widetilde{\mathbf{x}}^{(m)} (\widetilde{\mathbf{x}}^{(m)})^\top\right],$$

where $\delta(\cdot)$ is the Dirac delta function.

Convolution smoothing preserves convexity and provides a differentiable loss for the projected proximal-gradient algorithm described in the Supplementary Material. Although nonsmooth optimization methods can also address the original hinge-loss problem, smoothing makes gradient-based computation more convenient. Following the smoothing strategy used for single-task SVMs in Wang et al. (2023, 2024), we replace ϕ by its kernel-smoothed version

$$\phi_h(z) := (\phi * K_h)(z) = \int_{-\infty}^{\infty} \phi(z - v) K_h(v) dv, \qquad K_h(v) = h^{-1} K(v/h),$$

where K is a symmetric, non-negative kernel function satisfying $\int_{-\infty}^{\infty} K(u)du = 1$, and $h > 0$

Statistica Sinica: Newly accepted Paper

is a bandwidth parameter. For each task, define the margin variable

$$r_i^{(m)}(b^{(m)}, \boldsymbol{\mu}, \boldsymbol{\alpha}^{(m)}) := y_i^{(m)} \left((\mathbf{x}_i^{(m)})^\top (\boldsymbol{\mu} + \boldsymbol{\alpha}^{(m)}) + b^{(m)} \right).$$

Replacing ϕ by ϕ_h in (2.1), we obtain the multitask smoothed SVM estimator with composite regularization:

$$\begin{aligned} &(\widehat{\mathbf{b}}_h, \widehat{\boldsymbol{\mu}}_h, \widehat{\mathbf{A}}_h) \\ &:= \underset{\substack{\mathbf{b},\boldsymbol{\mu},\mathbf{A} \\ \mathbf{1}^\top \boldsymbol{\alpha}_j = 0, j \in [p]}}{\operatorname{argmin}} \left\{ \frac{1}{N} \sum_{m=1}^{M} \sum_{i=1}^{n_m} \phi_h \left(r_i^{(m)}(b^{(m)}, \boldsymbol{\mu}, \boldsymbol{\alpha}^{(m)}) \right) + \lambda \Big(\|\boldsymbol{\mu}\|_1 + \lambda_g \|\mathbf{A}\|_{2,1} \Big) \right\}. \end{aligned} \tag{2.3}$$

The above estimator can be numerically obtained via the proximal-gradient algorithm, which is detailed in the Supplementary Material.

To define the corresponding population target under heterogeneous tasks, let

$$R_h^{(m)}(b^{(m)}, \boldsymbol{\mu}, \boldsymbol{\alpha}^{(m)}) := \mathbb{E}^{(m)} \left[\phi_h \left(y^{(m)} \big((\mathbf{x}^{(m)})^\top (\boldsymbol{\mu} + \boldsymbol{\alpha}^{(m)}) + b^{(m)} \big) \right) \right].$$

We then define the population target by

$$(\mathbf{b}_h^*, \boldsymbol{\mu}_h^*, \mathbf{A}_h^*) \in \underset{\substack{\mathbf{b},\boldsymbol{\mu},\mathbf{A} \\ \mathbf{1}^\top \boldsymbol{\alpha}_j = 0, j \in [p]}}{\operatorname{argmin}} \frac{1}{N} \sum_{m=1}^{M} n_m R_h^{(m)}(b^{(m)}, \boldsymbol{\mu}, \boldsymbol{\alpha}^{(m)}). \tag{2.4}$$

Under suitable regularity conditions, the smoothed population minimizer is expected to approach $(\mathbf{b}^*, \boldsymbol{\mu}^*, \mathbf{A}^*)$ as $h \to 0$. In the statistical analysis below we measure the estimator relative to the nonsmoothed population target $(\mathbf{b}^*, \boldsymbol{\mu}^*, \mathbf{A}^*)$. The smoothing bias is therefore kept explicitly in the rates through a term $O(h^2)$, while the stochastic part is controlled by

Statistica Sinica: Newly accepted Paper

concentration over the multitask parameter blocks.

For each task m, define the task-specific score of the expected smoothed hinge loss with respect to the augmented parameter $\boldsymbol{\theta}^{(m)}$ by

$$\begin{aligned} & \mathbf{s}_h^{(m)}(\boldsymbol{\theta}^{(m)}) \\ := & \left(\mathbb{E}^{(m)}\left[\phi_h'\left(y^{(m)}(\widetilde{\mathbf{x}}^{(m)})^\top\boldsymbol{\theta}^{(m)}\right)y^{(m)}\right], \mathbb{E}^{(m)}\left[\phi_h'\left(y^{(m)}(\widetilde{\mathbf{x}}^{(m)})^\top\boldsymbol{\theta}^{(m)}\right)y^{(m)}(\mathbf{x}^{(m)})^\top\right]\right)^\top \\ = & -\mathbb{E}^{(m)}\left[\bar{K}\left(\frac{1-y^{(m)}(\widetilde{\mathbf{x}}^{(m)})^\top\boldsymbol{\theta}^{(m)}}{h}\right)y^{(m)}\widetilde{\mathbf{x}}^{(m)}\right]. \end{aligned}$$

Similarly, the corresponding task-specific Hessian matrix is

$$\begin{aligned} & \mathbf{H}_h^{(m)}(\boldsymbol{\theta}^{(m)}) \\ := & \begin{pmatrix} \mathbb{E}^{(m)}\left[\phi_h''\left(y^{(m)}(\widetilde{\mathbf{x}}^{(m)})^\top\boldsymbol{\theta}^{(m)}\right)\right] & \mathbb{E}^{(m)}\left[\phi_h''\left(y^{(m)}(\widetilde{\mathbf{x}}^{(m)})^\top\boldsymbol{\theta}^{(m)}\right)(\mathbf{x}^{(m)})^\top\right] \\ \mathbb{E}^{(m)}\left[\phi_h''\left(y^{(m)}(\widetilde{\mathbf{x}}^{(m)})^\top\boldsymbol{\theta}^{(m)}\right)\mathbf{x}^{(m)}\right] & \mathbb{E}^{(m)}\left[\phi_h''\left(y^{(m)}(\widetilde{\mathbf{x}}^{(m)})^\top\boldsymbol{\theta}^{(m)}\right)\mathbf{x}^{(m)}(\mathbf{x}^{(m)})^\top\right] \end{pmatrix} \\ = & \mathbb{E}^{(m)}\left[K_h\left(1-y^{(m)}(\widetilde{\mathbf{x}}^{(m)})^\top\boldsymbol{\theta}^{(m)}\right)\widetilde{\mathbf{x}}^{(m)}(\widetilde{\mathbf{x}}^{(m)})^\top\right], \end{aligned}$$

where $\phi_h'(u) = -\bar{K}(\frac{1-u}{h})$, $\bar{K}(z) := \int_{-\infty}^{z} K(t)dt$, and $\phi_h''(u) = K_h(1-u)$. Finally, we write the corresponding task-specific empirical score and Hessian matrix as

$$\begin{aligned} \widehat{\mathbf{s}}_h^{(m)}(\boldsymbol{\theta}^{(m)}) & := -\frac{1}{n_m}\sum_{i=1}^{n_m}\bar{K}\left(\frac{1-y_i^{(m)}(\widetilde{\mathbf{x}}_i^{(m)})^\top\boldsymbol{\theta}^{(m)}}{h}\right)y_i^{(m)}\widetilde{\mathbf{x}}_i^{(m)}, \\ \widehat{\mathbf{H}}_h^{(m)}(\boldsymbol{\theta}^{(m)}) & := \frac{1}{n_m}\sum_{i=1}^{n_m}K_h\left(1-y_i^{(m)}(\widetilde{\mathbf{x}}_i^{(m)})^\top\boldsymbol{\theta}^{(m)}\right)\widetilde{\mathbf{x}}_i^{(m)}(\widetilde{\mathbf{x}}_i^{(m)})^\top. \end{aligned}$$

3. Theoretical Analysis

In this section, we provide the main theoretical results on estimation error and variable selection of the proposed multitask estimator.

3.1 Estimation Error Bounds

To establish the estimation error bounds, we first impose the following conditions. **For a real-valued random variable Z, define its sub-Gaussian Orlicz norm by** $\|Z\|_{\psi_2} := \inf\{t > 0 : \mathbb{E}[\exp(Z^2/t^2)] \leq 2\}$.

Condition 1. The covariates are uniformly sub-Gaussian across tasks and coordinates in the sense that there exists a constant $K_x > 0$ such that

$$\sup_{1\leq m\leq M} \sup_{1\leq j\leq p} \|x_j^{(m)}\|_{\psi_2} \leq K_x.$$

Moreover, there exists some positive constant C such that

$$\max_{1\leq m\leq M} \lambda_{\max}\left(\mathbb{E}^{(m)}[\mathbf{x}^{(m)}(\mathbf{x}^{(m)})^\top]\right) \leq C.$$

Condition 2. Let $\boldsymbol{\beta}^{*(m)} := \boldsymbol{\mu}^* + \boldsymbol{\alpha}^{*(m)}$ for $m = 1, \ldots, M$. Assume that the shared component $\boldsymbol{\mu}^*$ is sparse with support $S_\mu := \{j : \mu_j^* \neq 0\}$ and $s_\mu := |S_\mu|$, and that the task-specific deviation matrix $\mathbf{A}^* = (\boldsymbol{\alpha}^{(1)*}, \ldots, \boldsymbol{\alpha}^{(M)*})$ is row-sparse with active row set $S_\alpha := \{j : \|\mathbf{A}_{j\cdot}^*\| \neq 0\}$ and $s_\alpha := |S_\alpha|$. Moreover, assume that $\boldsymbol{\beta}^{*(m)} \neq 0$. **For each task, choose** $j_m \in \arg\max_{1\leq j\leq p} |\beta_j^{*(m)}|$**. In the task-wise conditional-density calculations only, we label this coordinate as the first coordinate without loss of generality, writing** $\beta_1^{*(m)}$ **for**

Statistica Sinica: Newly accepted Paper

$\beta_{j_m}^{*(m)}$. **This local convention leaves the global feature indexing unchanged. Define** $\beta_{1,\min}^* := \min_{1\le m\le M} |\beta_{j_m}^{*(m)}| = \min_{1\le m\le M} \|\boldsymbol{\beta}^{*(m)}\|_\infty$**.** In addition, assume that $h = o((\beta_{1,\min}^*)^2)$ and $\sqrt{M}\lambda r_s = o(\beta_{1,\min}^*)$, where $r_s := \sqrt{s_\mu + s_\alpha}$.

Condition 3. For some sufficiently large constant $C_1 > 0$, we have

$$\inf_{\substack{\left(M^{-1}\|\mathbf{b}-\mathbf{b}^*\|^2+\|\boldsymbol{\mu}-\boldsymbol{\mu}^*\|^2+M^{-1}\|\mathbf{A}-\mathbf{A}^*\|_F^2\right)^{1/2}\le C_1\lambda r_s \\ (\mathbf{b}-\mathbf{b}^*,\boldsymbol{\mu}-\boldsymbol{\mu}^*,\mathbf{A}-\mathbf{A}^*)\in\mathbb{C}}} \min_{1\le m\le M} \lambda_{\min}\big(\mathbf{H}^{(m)}(\boldsymbol{\theta}^{(m)})\big) \ge C_2,$$

for some constant $C_2 > 0$, where $\boldsymbol{\theta}^{(m)} = \big(b^{(m)}, (\boldsymbol{\mu} + \boldsymbol{\alpha}^{(m)})^\top\big)^\top$, and

$$\mathbb{C} := \Big\{(\mathbf{u},\mathbf{v},\boldsymbol{\Delta}) \in \mathbb{R}^M \times \mathbb{R}^p \times \mathbb{R}^{p\times M} : \\ \|\mathbf{v}_{S_\mu^c}\|_1 + \lambda_g\|\boldsymbol{\Delta}_{S_\alpha^c}\|_{2,1} \le 3\big(\|\mathbf{v}_{S_\mu}\|_1 + \lambda_g\|\boldsymbol{\Delta}_{S_\alpha}\|_{2,1}\big) + M^{-1/2}\|\mathbf{u}\|_2\Big\},$$

where $\boldsymbol{\Delta}_{S_\alpha}$, for example, denotes the submatrix of $\boldsymbol{\Delta}$ consisting of the rows indexed by S_α.

For each task $m \in [M]$, let $f^{(m)}$ and $g^{(m)}$ denote the conditional densities of $\mathbf{x}^{(m)}$ given $y^{(m)} = 1$ and $y^{(m)} = -1$, respectively. **In the local coordinate convention of Condition 2,** $x_1^{(m)}$ **denotes** $x_{j_m}^{(m)}$ **and** $\mathbf{x}_{-1}^{(m)}$ **contains the remaining coordinates in their original relative order.** Let $f_{-1}^{(m)}(\mathbf{x}_{-1})$ denote the joint density of $\mathbf{x}_{-1}^{(m)}$ conditional on $y^{(m)} = 1$, and define $g_{-1}^{(m)}$ analogously. **Condition 4 below is imposed on the conditional density of this selected coordinate given the remaining coordinates, with constants uniform across tasks.**

Condition 4. Assume that, for each m, $f^{(m)}$ and $g^{(m)}$ are continuously differentiable. In addition, for any fixed $\boldsymbol{\alpha} \in \mathbb{R}^{p-1}$ with $\|\boldsymbol{\alpha}\| = 1$, there exists a function $\mathcal{E}^{(m)}(\mathbf{x}_{-1})$ such that

Statistica Sinica: Newly accepted Paper

$\int_{\mathbb{R}^{p-1}} \mathcal{E}^{(m)}(\mathbf{x}_{-1}) f_{-1}^{(m)}(\mathbf{x}_{-1}) d\mathbf{x}_{-1} \le C$, and

$$\sup_x \big\{(1+|x|+|\boldsymbol{\alpha}^\top \mathbf{x}_{-1}|) f^{(m)}(x|\mathbf{x}_{-1}) + (1+|x|+|\boldsymbol{\alpha}^\top \mathbf{x}_{-1}|+|x|^2+|\boldsymbol{\alpha}^\top \mathbf{x}_{-1}|^2)|(f^{(m)})'(x|\mathbf{x}_{-1})|\big\} \le \mathcal{E}^{(m)}(\mathbf{x}_{-1}),$$

with the same assumption imposed on $g^{(m)}$. **The envelope may depend on the direction $\boldsymbol{\alpha}$, while its integral bound is uniform over unit directions.**

Condition 5. The kernel function $K : \mathbb{R} \to [0, +\infty)$ satisfies $K(u) = K(-u)$, $\int_{\mathbb{R}} K(u)du = 1$ and $\int_{\mathbb{R}} u^2 K(u)du \le C$.

Remark 1. Conditions 1–5 impose standard local regularity requirements for high-dimensional smoothed margin estimation. Condition 1 can be satisfied by Gaussian, bounded, and many other light-tailed covariate distributions. Condition 2 encodes the modeling assumption that most features are either inactive or homogeneous across tasks. Condition 3 is a restricted curvature condition, analogous to a restricted eigenvalue condition, and is natural when the class-conditional distributions place sufficient mass near the decision boundary in sparse directions. Condition 4 is **a sufficient technical condition for controlling the smoothing bias and Hessian approximation; we do not claim it is necessary. It holds for class-conditional Gaussian covariates whose covariance eigenvalues lie between fixed positive constants and whose mean vectors have uniformly bounded Euclidean norms, across tasks and classes. A derivation is given in the final subsection of the Supplementary Material, "Discussion and Gaussian verification of Condition 4." The density condition concerns only the random covariates: the augmented intercept coordinate is the constant 1 and has no Lebesgue density.**

Statistica Sinica: Newly accepted Paper

Condition 5 holds for standard symmetric smoothing kernels with finite second moment. **The coefficient dependence arises from controlling smoothing near the margin boundary. Conditional on the remaining covariates, a margin width h corresponds to a width $h/|\beta^{*(m)}_{j_m}|$ along coordinate j_m. The change of variables in the supplementary proofs therefore yields score bias of order $h^2/|\beta^{*(m)}_{j_m}|^2$ and Hessian approximation error of order $h/|\beta^{*(m)}_{j_m}|^2$. Choosing a largest-magnitude coordinate gives the strongest of these coordinatewise bounds.**

Theorem 1. Assume that Conditions 1–5 hold, and that the task sample sizes are balanced in the sense that

$$\frac{c_-}{M} \le \frac{n_m}{N} \le \frac{c_+}{M}, \qquad m \in [M],$$

for some constants $0 < c_- < c_+ < \infty$. Suppose further that $\log^3(p \vee N) = O(N)$. Define $r_s = \sqrt{s_\mu + s_\alpha}$, and choose the tuning parameters such that $\lambda_g \asymp M^{-1/2}$ and

$$\lambda \ge C \left\{ \sqrt{\frac{M + \log(p \vee N)}{N}} + \left(\frac{h}{|\beta^*_{1,\min}|} \right)^2 \right\},$$

for a sufficiently large constant $C > 0$. Then, with probability at least $1 - (p \vee N)^{-C}$, the estimator $(\widehat{\mathbf{b}}_h, \widehat{\boldsymbol{\mu}}_h, \widehat{\mathbf{A}}_h)$ satisfies

$$\frac{1}{\sqrt{M}} \|\widehat{\mathbf{b}}_h - \mathbf{b}^*\| + \|\widehat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}^*\| + \frac{1}{\sqrt{M}} \|\widehat{\mathbf{A}}_h - \mathbf{A}^*\|_F \lesssim r_s \lambda.$$

Theorem 1 shows that the estimation error depends on the effective sparsity through $r_s = \sqrt{s_\mu + s_\alpha}$, rather than directly on the ambient dimension p. The stochastic term

Statistica Sinica: Newly accepted Paper

$\sqrt{\{M + \log(p \vee N)\}/N}$ captures the cost of simultaneously estimating the task-specific components and controlling the fluctuations over the high-dimensional feature space. The term $(h/|\beta^*_{1,\min}|)^2$ quantifies the approximation bias induced by smoothing the hinge loss.

For bandwidth selection, let $a_N = \sqrt{\{M + \log(p \vee N)\}/N}$. Choosing $\lambda \asymp a_N + (h/\beta^*_{1,\min})^2$, with a sufficiently large constant, gives a normalized estimation error of order $r_s\{a_N + (h/\beta^*_{1,\min})^2\}$ in Theorem 1. If $(h/\beta^*_{1,\min})^2 = O(a_N)$, equivalently

$$h = O\left(\beta^*_{1,\min}\left\{\frac{M + \log(p \vee N)}{N}\right\}^{1/4}\right),$$

the bound has order $r_s a_N$, provided the remaining assumptions, including $h = o((\beta^*_{1,\min})^2)$ and $\sqrt{M}\lambda r_s = o(\beta^*_{1,\min})$, hold. The stochastic term is independent of h, so the bound does not identify a unique optimal bandwidth.

Remark 2. Let $\boldsymbol{\delta}_{\mathbf{b}} = \widehat{\mathbf{b}}_h - \mathbf{b}^*$, $\boldsymbol{\delta}_{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}_h - \boldsymbol{\mu}^*$, and $\boldsymbol{\Delta}_{\mathbf{A}} = \widehat{\mathbf{A}}_h - \mathbf{A}^*$. The centering constraint gives $\sum_{m=1}^M \boldsymbol{\Delta}_{\mathbf{A}}^{(m)} = 0$, and hence

$$\frac{1}{M}\sum_{m=1}^{M} \|\boldsymbol{\delta}_{\boldsymbol{\mu}} + \boldsymbol{\Delta}_{\mathbf{A}}^{(m)}\|_2^2 = \|\boldsymbol{\delta}_{\boldsymbol{\mu}}\|_2^2 + \frac{1}{M}\|\boldsymbol{\Delta}_{\mathbf{A}}\|_F^2.$$

Thus, the $M^{-1/2}$ factors measure root mean squared errors across tasks, while the common component remains unscaled. Under balanced task sizes, this normalization agrees with the sample-weighted quadratic error and the neighborhood in Condition 3. It also matches the cone's intercept term $M^{-1/2}\|\mathbf{u}\|_2$ and the group-penalty scaling $\lambda_g \asymp M^{-1/2}$. On the same high-probability event, Theorem 1 implies

$$\|\boldsymbol{\delta}_{\boldsymbol{\mu}}\|_2 \lesssim r_s\lambda, \qquad \|\boldsymbol{\delta}_{\mathbf{b}}\|_2 \lesssim \sqrt{M}\, r_s\lambda, \qquad \|\boldsymbol{\Delta}_{\mathbf{A}}\|_F \lesssim \sqrt{M}\, r_s\lambda.$$

Statistica Sinica: Newly accepted Paper

Establishing sharper component-specific rates requires further analytical work and remains open in the present framework.

Remark 3. Theorem 1 allows M to grow with N, provided the regularity constants are uniform and all stated conditions hold. If the smoothing bias is no larger than the stochastic term, choose $\lambda \asymp \sqrt{\{M + \log(p \vee N)\}/N}$. The requirement $\sqrt{M}\lambda r_s = o(\beta^*_{1,\min})$ then becomes

$$M(s_\mu + s_\alpha)\{M + \log(p \vee N)\} = o\left(N(\beta^*_{1,\min})^2\right).$$

For fixed sparsity and $\beta^*_{1,\min}$ bounded away from zero, it suffices that $M^2 = o(N)$ and $M\log(p \vee N) = o(N)$. In particular, when $\log(p \vee N) = O(M)$, $M = o(\sqrt{N})$ suffices for this requirement.

3.2 Variable Selection Consistency

In this subsection, we establish variable selection consistency of the multitask smoothed SVM estimator. Recall that the task-specific augmented parameter for task m is $\boldsymbol{\theta}^{(m)} = \left(b^{(m)}, (\boldsymbol{\mu} + \boldsymbol{\alpha}^{(m)})^\top\right)^\top \in \mathbb{R}^{p+1}$. To accommodate the identifiability constraints $\mathbf{1}^\top \boldsymbol{\alpha}_j = 0$ for all $j \in [p]$, we eliminate $\boldsymbol{\alpha}^{(1)}$ by writing $\boldsymbol{\alpha}^{(1)} = -\sum_{m=2}^{M} \boldsymbol{\alpha}^{(m)}$. We then define the global reparameterized vector

$$\boldsymbol{\vartheta} = \left(\mathbf{b}^\top, \boldsymbol{\mu}^\top, (\boldsymbol{\alpha}^{(2)})^\top, \ldots, (\boldsymbol{\alpha}^{(M)})^\top\right)^\top \in \mathbb{R}^{M(p+1)}.$$

The task-specific parameters can be recovered from $\boldsymbol{\vartheta}$ through $\boldsymbol{\theta}^{(1)}(\boldsymbol{\vartheta}) = (b^{(1)}, \{\boldsymbol{\mu} - \sum_{k=2}^{M} \boldsymbol{\alpha}^{(k)}\}^\top)^\top$ and $\boldsymbol{\theta}^{(m)}(\boldsymbol{\vartheta}) = \left(b^{(m)}, (\boldsymbol{\mu} + \boldsymbol{\alpha}^{(m)})^\top\right)^\top$ for $m = 2, \ldots, M$. Accordingly, the smoothed empirical

Statistica Sinica: Newly accepted Paper

loss, viewed as a function of $\boldsymbol{\vartheta}$, is given by

$$\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}) = \frac{1}{N}\sum_{m=1}^{M}\sum_{i=1}^{n_m}\phi_h\left(y_i^{(m)}(\widetilde{\mathbf{x}}_i^{(m)})^\top\boldsymbol{\theta}^{(m)}(\boldsymbol{\vartheta})\right).$$

For each $j \in [p]$, define

$$\mathbf{a}_j := (\alpha_j^{(2)}, \ldots, \alpha_j^{(M)})^\top \in \mathbb{R}^{M-1} \quad \text{and} \quad \mathbf{T} = \begin{pmatrix} -\mathbf{1}_{M-1}^\top \\ \mathbb{I}_{M-1} \end{pmatrix} \in \mathbb{R}^{M\times(M-1)}.$$

Then $\mathbf{T}\mathbf{a}_j = (\alpha_j^{(1)}, \alpha_j^{(2)}, \ldots, \alpha_j^{(M)})^\top = \mathbf{A}_{j\cdot}^\top$, and hence $\|\mathbf{T}\mathbf{a}_j\|_2 = \|\mathbf{A}_{j\cdot}\|_2$. Therefore, the composite penalty can be expressed in terms of the reparameterized vector as

$$\mathcal{P}(\boldsymbol{\vartheta}) = \|\boldsymbol{\mu}\|_1 + \lambda_g\sum_{j=1}^{p}\|\mathbf{T}\mathbf{a}_j\|_2.$$

This reparameterization yields an unconstrained optimization problem in $\mathbf{a}_j \in \mathbb{R}^{M-1}$ while preserving the original group penalty on the centered task-specific deviations. Accordingly, the multitask smoothed SVM estimator is equivalently given by

$$\widehat{\boldsymbol{\vartheta}}_h = \underset{\boldsymbol{\vartheta}}{\operatorname{argmin}}\left\{\widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta}) + \lambda\mathcal{P}(\boldsymbol{\vartheta})\right\}.$$

Let

$$\boldsymbol{\vartheta}^* = \left(\mathbf{b}^{*\top}, \boldsymbol{\mu}^{*\top}, \boldsymbol{\alpha}^{*(2)\top}, \ldots, \boldsymbol{\alpha}^{*(M)\top}\right)^\top \in \mathbb{R}^{M(p+1)}$$

be the reparameterized form of the original population target. Define the estimated active

Statistica Sinica: Newly accepted Paper

sets

$$\widehat{S}_\mu = \{j : \widehat{\mu}_{h,j} \neq 0\}, \qquad \widehat{S}_\alpha = \{j : \|\widehat{\mathbf{A}}_{h,j\cdot}\|_2 \neq 0\}.$$

Define the active block as

$$\mathcal{I} = \{\text{all intercept coordinates}\} \cup \{\mu_j : j \in S_\mu\} \cup \{\mathbf{a}_j : j \in S_\alpha\},$$

whereas the inactive block is

$$\mathcal{I}^c = \left\{\mu_j : j \in S_\mu^c\right\} \cup \left\{\mathbf{a}_j : j \in S_\alpha^c\right\}.$$

Note that all intercept coordinates are included in the active block because they are not penalized. For any vector $\mathbf{u}$ partitioned as

$$\mathbf{u} = (\mathbf{u}_\mathbf{b}^\top, \mathbf{u}_{\boldsymbol{\mu}}^\top, \mathbf{u}_\mathbf{a}^\top)^\top \in \mathbb{R}^{M(p+1)},$$

where $\mathbf{u}_\mathbf{b} \in \mathbb{R}^M$, $\mathbf{u}_{\boldsymbol{\mu}} \in \mathbb{R}^p$ and $\mathbf{u}_\mathbf{a} = (\mathbf{u}_{\mathbf{a},1}^\top, \ldots, \mathbf{u}_{\mathbf{a},p}^\top)^\top \in \mathbb{R}^{p(M-1)}$ with

$$\mathbf{u}_{\mathbf{a},j} = \left(u_{a,j}^{(2)}, u_{a,j}^{(3)}, \ldots, u_{a,j}^{(M)}\right)^\top \in \mathbb{R}^{M-1} \quad j = 1, \ldots, p,$$

we write $\mathbf{u}_\mathcal{I}$ for the subvector of $\mathbf{u}$ restricted to the active block $\mathcal{I}$. More precisely, if $S_\alpha = \{j_1, \ldots, j_{s_\alpha}\}$ with the indices arranged in increasing order, define

$$\mathbf{u}_{\mathbf{a},S_\alpha} = \left(\mathbf{u}_{\mathbf{a},j_1}^\top, \ldots, \mathbf{u}_{\mathbf{a},j_{s_\alpha}}^\top\right)^\top \in \mathbb{R}^{(M-1)s_\alpha}.$$

Statistica Sinica: Newly accepted Paper

Then

$$\mathbf{u}_{\mathcal{I}} = \left(\mathbf{u}_{\mathbf{b}}^{\top}, \mathbf{u}_{\boldsymbol{\mu},S_\mu}^{\top}, \mathbf{u}_{\mathbf{a},S_\alpha}^{\top}\right)^{\top} \in \mathbb{R}^{M+s_\mu+(M-1)s_\alpha}.$$

We next introduce norms adapted to the geometry of the composite penalty. For any $\mathbf{a} \in \mathbb{R}^{M-1}$, define $\|\mathbf{a}\|_{\mathbf{T}} = \|\mathbf{T}\mathbf{a}\|_2$ and the dual norm by $\|\mathbf{z}\|_{\widetilde{\mathbf{T}}} = \left\{\mathbf{z}^{\top}(\mathbf{T}^{\top}\mathbf{T})^{-1}\mathbf{z}\right\}^{1/2}$, or equivalently, $\|\mathbf{z}\|_{\widetilde{\mathbf{T}}} = \left\|\mathbf{T}(\mathbf{T}^{\top}\mathbf{T})^{-1}\mathbf{z}\right\|_2$. For the penalized coordinates, define the mixed dual norm

$$\|\mathbf{u}\|_{\mathcal{D}} := \max\left\{\|\mathbf{u}_{\boldsymbol{\mu}}\|_\infty,\ \lambda_g^{-1}\max_{1\le j\le p}\|\mathbf{u}_{\mathbf{a},j}\|_{\widetilde{\mathbf{T}}}\right\}.$$

For a vector involving the intercepts, we use the augmented mixed norm

$$\|\mathbf{u}\|_{\mathcal{D},+} := \max\left\{M^{-1/2}\|\mathbf{u}_{\mathbf{b}}\|_2,\ \|\mathbf{u}_{\boldsymbol{\mu}}\|_\infty,\ \lambda_g^{-1}\max_{1\le j\le p}\|\mathbf{u}_{\mathbf{a},j}\|_{\widetilde{\mathbf{T}}}\right\}.$$

When this norm is applied to an active-block vector $\mathbf{u}_{\mathcal{I}}$, it is understood as

$$\|\mathbf{u}_{\mathcal{I}}\|_{\mathcal{D},+} = \max\left\{M^{-1/2}\|\mathbf{u}_{\mathbf{b}}\|_2,\ \|\mathbf{u}_{\boldsymbol{\mu},S_\mu}\|_\infty,\ \lambda_g^{-1}\max_{j\in S_\alpha}\|\mathbf{u}_{\mathbf{a},j}\|_{\widetilde{\mathbf{T}}}\right\}.$$

Similarly, for an inactive-block vector,

$$\|\mathbf{u}_{\mathcal{I}^c}\|_{\mathcal{D}} = \max\left\{\|\mathbf{u}_{\boldsymbol{\mu},S_\mu^c}\|_\infty,\ \lambda_g^{-1}\max_{j\in S_\alpha^c}\|\mathbf{u}_{\mathbf{a},j}\|_{\widetilde{\mathbf{T}}}\right\}.$$

For any differentiable function $G(\boldsymbol{\vartheta})$, we use $\nabla_{\mathcal{I}}G(\boldsymbol{\vartheta})$ to denote the subvector of the full gradient $\nabla_{\boldsymbol{\vartheta}}G(\boldsymbol{\vartheta})$ corresponding to the active block $\mathcal{I}$. Specifically, if $S_\alpha = \{j_1, \ldots, j_{s_\alpha}\}$

Statistica Sinica: Newly accepted Paper

with the indices arranged in increasing order, then

$$\nabla_{\mathcal{I}} G(\boldsymbol{\vartheta}) = \left(\nabla_{\mathbf{b}} G(\boldsymbol{\vartheta})^\top, \{\nabla_{\boldsymbol{\mu}} G(\boldsymbol{\vartheta})\}_{S_\mu}^\top, \nabla_{\mathbf{a}_{j_1}} G(\boldsymbol{\vartheta})^\top, \ldots, \nabla_{\mathbf{a}_{j_{s_\alpha}}} G(\boldsymbol{\vartheta})^\top \right)^\top.$$

Here $\nabla_{\mathbf{a}_j} G(\boldsymbol{\vartheta}) \in \mathbb{R}^{M-1}$ denotes the gradient with respect to $\mathbf{a}_j = (\alpha_j^{(2)}, \ldots, \alpha_j^{(M)})^\top$. Similarly, if $S_\alpha^c = \{k_1, \ldots, k_{p-s_\alpha}\}$, then

$$\nabla_{\mathcal{I}^c} G(\boldsymbol{\vartheta}) = \left(\{\nabla_{\boldsymbol{\mu}} G(\boldsymbol{\vartheta})\}_{S_\mu^c}^\top, \nabla_{\mathbf{a}_{k_1}} G(\boldsymbol{\vartheta})^\top, \ldots, \nabla_{\mathbf{a}_{k_{p-s_\alpha}}} G(\boldsymbol{\vartheta})^\top \right)^\top.$$

Finally, let

$$\boldsymbol{\Gamma}_h(\boldsymbol{\vartheta}) = \nabla^2_{\boldsymbol{\vartheta}} \widetilde{\mathcal{L}}_{N,h}(\boldsymbol{\vartheta})$$

denote the empirical Hessian matrix of the reparameterized smoothed loss. For two index sets or block index sets A and B, we use $\boldsymbol{\Gamma}_{h,AB}(\boldsymbol{\vartheta})$ to denote the submatrix of $\boldsymbol{\Gamma}_h(\boldsymbol{\vartheta})$ with rows indexed by A and columns indexed by B. In particular, $\boldsymbol{\Gamma}_{h,\mathcal{I}\mathcal{I}}(\boldsymbol{\vartheta}^*)$ is the active block empirical Hessian, and $\boldsymbol{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*)$ denotes the empirical Hessian block with rows in the inactive block $\mathcal{I}^c$ and columns in the active block $\mathcal{I}$.

We impose the following additional conditions for variable selection consistency.

Condition 6. There exists a constant $c_{\min} > 0$ such that, with probability tending to one,

$$\lambda_{\min}\left(\boldsymbol{\Gamma}_{h,\mathcal{I}\mathcal{I}}(\boldsymbol{\vartheta}^*)\right) \geq c_{\min}.$$

Condition 6 requires nonsingularity only on the active block, whose dimension is $M +$

Statistica Sinica: Newly accepted Paper

$s_\mu + (M-1)s_\alpha$. It is therefore weaker than requiring the full empirical Hessian in $M(p+1)$ dimensions to be well conditioned.

Condition 7. There exists some $\eta_0 \in (0,1)$ such that, with probability tending to one, for all active subgradients $\boldsymbol{\zeta}_{\mathcal{I}}$ satisfying $\boldsymbol{\zeta}_b = 0$, $|\boldsymbol{\zeta}_{\mu,j}| \le 1, j \in S_\mu$ and $\|\boldsymbol{\zeta}_{a,j}\|_{\widetilde{\mathbf{T}}} \le \lambda_g$ for $j \in S_\alpha$, we have

$$\left\| \left[\boldsymbol{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*) \boldsymbol{\Gamma}^{-1}_{h,\mathcal{I}\mathcal{I}}(\boldsymbol{\vartheta}^*) \boldsymbol{\zeta}_{\mathcal{I}} \right]_{\mu, S_\mu^c} \right\|_\infty \le 1 - \eta_0,$$

and

$$\max_{j \in S_\alpha^c} \left\| \left[\boldsymbol{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*) \boldsymbol{\Gamma}^{-1}_{h,\mathcal{I}\mathcal{I}}(\boldsymbol{\vartheta}^*) \boldsymbol{\zeta}_{\mathcal{I}} \right]_{a,j} \right\|_{\widetilde{\mathbf{T}}} \le \lambda_g (1 - \eta_0).$$

In addition, for the unpenalized intercept block, assume that there exists a constant $C_b > 0$ such that, with probability tending to one,

$$\left\| \boldsymbol{\Gamma}_{h,\mathcal{I}^c\mathcal{I}}(\boldsymbol{\vartheta}^*) \boldsymbol{\Gamma}^{-1}_{h,\mathcal{I}\mathcal{I}}(\boldsymbol{\vartheta}^*) \mathbf{u}_{\mathcal{I},b} \right\|_{\mathcal{D}} \le C_b M^{-1/2} \|\mathbf{u}_b\|_2,$$

for all vectors $\mathbf{u}_{\mathcal{I},b}$ supported only on the intercept block.

Condition 7 is the usual irrepresentable condition adapted to the composite $\ell_1/\ell_{2,1}$ penalty. It says that inactive features cannot be too well explained by active features through the local Hessian.

Define the minimum signal strength by

$$\nu_N = \min \left\{ \min_{j \in S_\mu} |\mu_j^*|,\ M^{-1/2} \min_{j \in S_\alpha} \|\mathbf{A}_{j\cdot}^*\|_2 \right\}.$$

Statistica Sinica: Newly accepted Paper

Condition 8. The minimum signal strength satisfies $r_s\lambda = o(\nu_N)$.

Condition 8 is not needed for estimation consistency, but is necessary for exact support recovery because signals smaller than the estimation error cannot be distinguished from zero. The scaling $M^{-1/2}\|\mathbf{A}^*_{j\cdot}\|_2$ puts the group signal for task-specific deviations on the same scale as the coordinatewise signal in $\boldsymbol{\mu}^*$.

We now state the variable selection consistency theorem.

Theorem 2. Assume that Conditions 1–8 hold, and that the task sample sizes are balanced in the sense that

$$\frac{c_-}{M} \le \frac{n_m}{N} \le \frac{c_+}{M}, \qquad m = 1, \dots, M,$$

for some constants $0 < c_- < c_+ < \infty$. Suppose further that the smoothing kernel is continuously differentiable with $\|K'\|_\infty < \infty$, $\lambda_g \asymp M^{-1/2}$ and $d_N = o(\eta_0\lambda)$, where

$$d_N = \sqrt{\frac{M + \log(p \vee N)}{N}} + \left(\frac{h}{|\beta^*_{1,\min}|}\right)^2.$$

Let $B_N = C\sqrt{\log(p \vee N)}$, $q_N = M + s_\mu + (M-1)s_\alpha$, and assume the bandwidth condition $B_N^3 M^2 q_N r_s^2 \lambda = o(h)$. Then we have

$$P\left(\widehat{S}_\mu = S_\mu,\ \widehat{S}_\alpha = S_\alpha\right) \to 1.$$

Theorem 2 establishes that, under additional irrepresentability and beta-min conditions, the proposed estimator consistently recovers both the support of the common component and the row support of the task-specific deviation matrix. It therefore distinguishes between the

two sources of sparsity: S_μ identifies features whose effects are shared across tasks, whereas S_α identifies features that contribute to task heterogeneity.

For support recovery, Theorem 2 additionally requires $B_N^3 M^2 q_N r_s^2 \lambda = o(h)$, so the bandwidth cannot be arbitrarily small relative to the regularization level. Smoothing may not be essential for an estimation bound analogous to Theorem 1, whose argument mainly uses convexity, Lipschitz continuity, and local population curvature, although we do not establish this extension here. Our proof of Theorem 2, however, explicitly uses the second derivative of the empirical loss: the empirical Hessian linearizes the score, its active-block inverse expresses the restricted estimator's error, and its inactive–active block is used to verify strict dual feasibility. The Taylor-remainder lemma in the Supplementary Material controls the Taylor remainder using the third derivative of the smoothed loss. Whether smoothing can be removed from the support-recovery analysis through a different proof technique remains open.

4. Simulations

We present simulation results in this section to evaluate the finite-sample performance of the proposed multitask smoothed SVM estimator with composite regularization. We generate binary classification data from a two-class Gaussian model. For each task $m = 1, \ldots, M$, the class label $y_i^{(m)}$ is generated according to $\mathbb{P}(y_i^{(m)} = 1) = \mathbb{P}(y_i^{(m)} = -1) = 0.5$. If $y_i^{(m)} = 1$, the predictor $\mathbf{x}_i^{(m)}$ is drawn from the multivariate normal distribution $\mathcal{N}(\boldsymbol{\delta}_m, \mathbf{I}_p)$; if $y_i^{(m)} = -1$, $\mathbf{x}_i^{(m)}$ is drawn from $\mathcal{N}(-\boldsymbol{\delta}_m, \mathbf{I}_p)$.To specify $\boldsymbol{\delta}_m$, we first construct the target population SVM

coefficient

$$\boldsymbol{\beta}^{*(m)} = \boldsymbol{\mu}^* + \boldsymbol{\alpha}^{*(m)}, \qquad m = 1, \ldots, M,$$

and then choose $\boldsymbol{\delta}_m$ such that the population SVM coefficient induced by the Gaussian model coincides with $\boldsymbol{\beta}^{*(m)}$. Specifically, we set $\boldsymbol{\mu}^* = (1.0, 0.6, 0.4, 0, \ldots, 0)^\top \in \mathbb{R}^p$. To generate task heterogeneity, we begin with a vector containing equal numbers of -1 and 1, add independent zero-mean Gaussian perturbations, and then center and rescale the resulting vector to obtain the vector, say, $(z_1, \ldots, z_M)^\top$, satisfying $\sum_{m=1}^M z_m = 0$ and $\sum_{m=1}^M z_m^2 = M$. We define the task-specific deviations by

$$\boldsymbol{\alpha}^{*(m)} = (0, 0, 0, az_m, -az_m, 0, \ldots, 0)^\top, \qquad m = 1, \ldots, M,$$

where $a > 0$ controls the degree of task heterogeneity. Therefore, the shared active set is $S_\mu = \{1, 2, 3\}$, and the row support of the task-specific deviation matrix is $S_\alpha = \{4, 5\}$. According to the population characterization of the linear SVM under the two-class Gaussian model (Koo et al., 2008), $\boldsymbol{\delta}_m$ and $\boldsymbol{\beta}^{*(m)}$ are related by

$$\boldsymbol{\beta}^{*(m)} = \frac{\boldsymbol{\delta}_m}{b_m \|\boldsymbol{\delta}_m\| + \|\boldsymbol{\delta}_m\|^2},$$

where b_m is determined by $\gamma(b_m) = \|\boldsymbol{\delta}_m\| \Phi(b_m)$. Here $\gamma(\cdot)$ and $\Phi(\cdot)$ denote the density and cumulative distribution functions of the standard normal distribution, respectively. $\boldsymbol{\delta}_m$ can then be obtained by solving the above displayed equation (details omitted).

Unless otherwise stated, we take $M = 10$, $p = 500$, $a = 0.5$, and use a Gaussian kernel with smoothing bandwidth $h = 0.2$ **(sensitivity to the choice of h is assessed**

in the Supplementary Material). For each task, the training and test sample sizes are $n_{\text{train}} = 200$ and $n_{\text{test}} = 200$, respectively. All results are averaged over 100 independent replications. We choose regularization parameters λ and λ_g through cross-validation. We evaluate classification and estimation performance using the misclassification rate (MR) and the estimation error (EE). We also report the true positive rate (TPR) and false positive rate (FPR) for recovering the active coefficients. The proposed multitask smoothed SVM is compared with the following two benchmark methods.

- Pooled: The estimator that pools all individual data across tasks and fits a single ℓ_1-penalized smoothed SVM (ignoring heterogeneity).
- Local: Fits a separate ℓ_1-penalized smoothed SVM to each task independently, using task-specific tuning parameters selected by cross-validation.

As practical tuning guidance, the bandwidth conditions involving unknown population quantities are sufficient theoretical conditions rather than directly implementable selection rules. Smaller h reduces smoothing bias but makes optimization and derivative-based analysis more demanding. We recommend comparing a finite grid of positive bandwidths, tuning the penalties for each candidate. Prediction-based cross-validation does not itself guarantee support recovery.

For practical implementation, we use five-fold cross-validation with folds formed within each task and stratified by class. For each candidate h, jointly search positive log-spaced grids for λ and $c_g = \sqrt{M}\lambda_g$, setting $\lambda_g = c_g/\sqrt{M}$. Use identical folds for all candidates and minimize the misclassification rate aggregated over all held-out observations, consistent with the sample-weighted ob-

jective. Refit on all training observations, and reserve the test data for final evaluation.

In the first simulation experiment, we study how the degree of task heterogeneity affects the performance of different methods. Specifically, we vary $a \in \{0, 0.3, 0.5, 0.7\}$. Figure 1 summarizes the results. When $a = 0$, all tasks share the same population SVM coefficient, so the pooled estimator is the best because it effectively uses all observations to estimate a single classifier. However, as heterogeneity increases (that is, a increases), the pooled estimator suffers from a non-negligible bias. The local estimator is less biased under heterogeneity but has a larger variance because each task uses only its own sample. The proposed multitask estimator provides a compromise between these two extremes. It borrows information through the common component $\boldsymbol{\mu}$ while allowing sparse task-specific deviations through $\mathbf{A}$. Consequently, it achieves smaller estimation error and misclassification rate when the tasks are related but not identical. The TPR/FPR remains high/low for the proposed method, suggesting that the composite penalty is effective in controlling false discoveries in high dimensions.

In the second simulation, we consider the effect of dimension $p \in \{300, 500, 700, 900\}$. As shown in Figure 2, methods generally become less accurate as p increases, especially for the local estimator, which has to estimate a high-dimensional classifier using only task-specific data.

In the third simulation, we vary the per-task sample size over $n \in \{200, 400, 600, 800\}$. Figure 3 shows that the estimation error and misclassification rate decrease as the sample size increases. The local method improves with larger n, but it remains less efficient since each task has a limited sample size. The proposed multitask estimator consistently yields

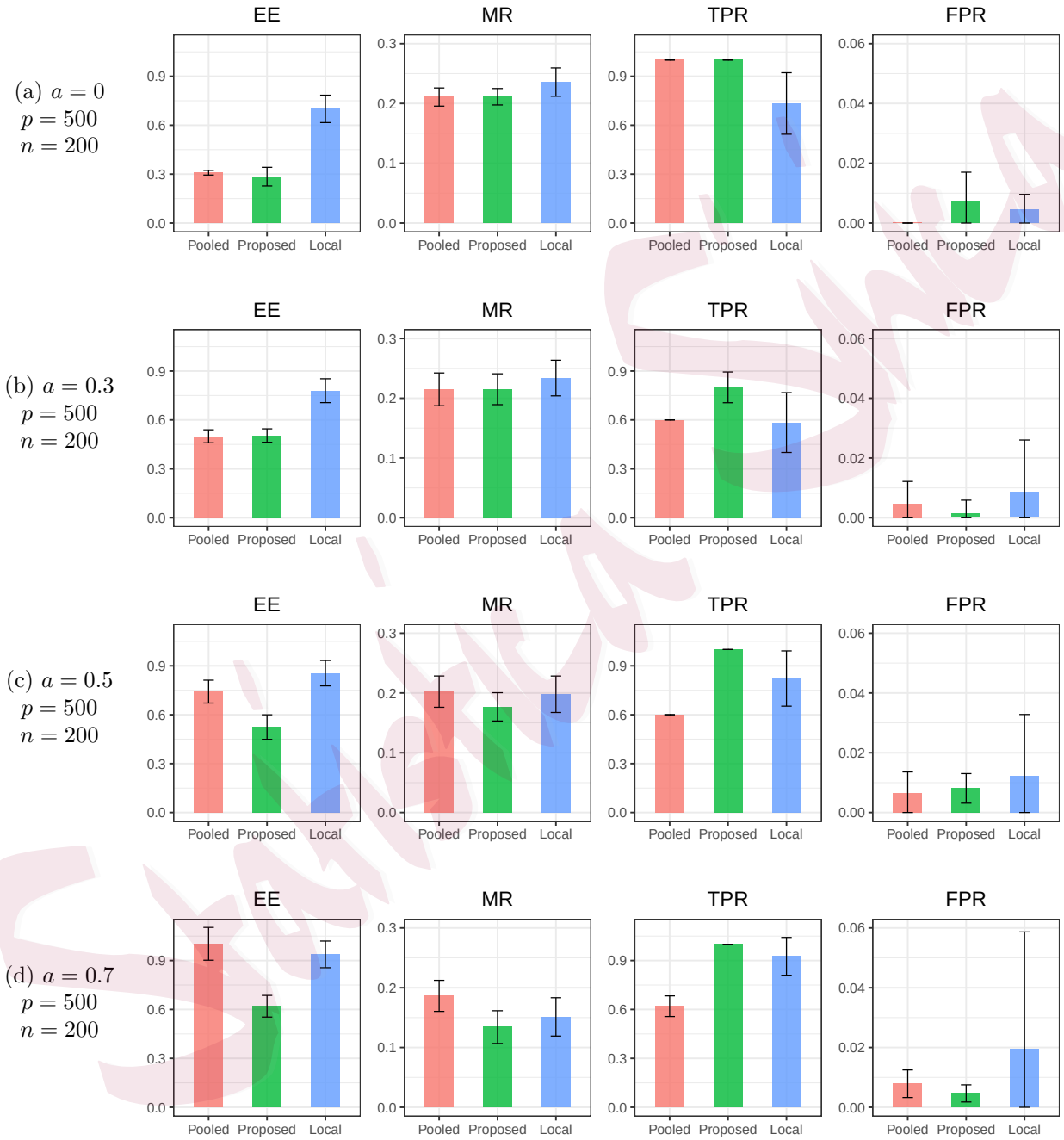


Figure 1: Comparison of different methods for varying $a \in \{0, 0.3, 0.5, 0.7\}$.

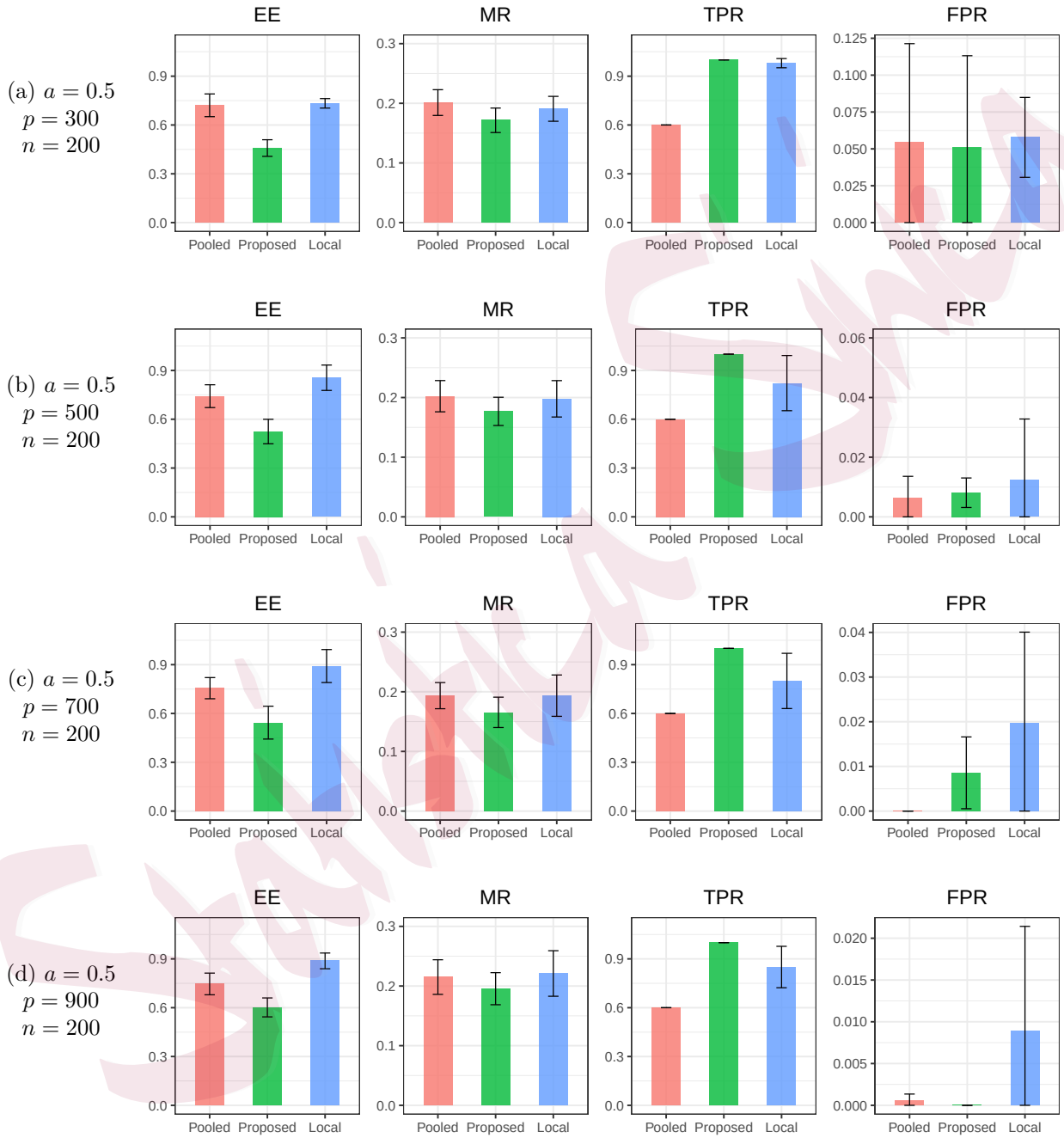


Figure 2: Comparison of different methods for varying $p \in \{300, 500, 700, 900\}$.

favorable performance across sample sizes.

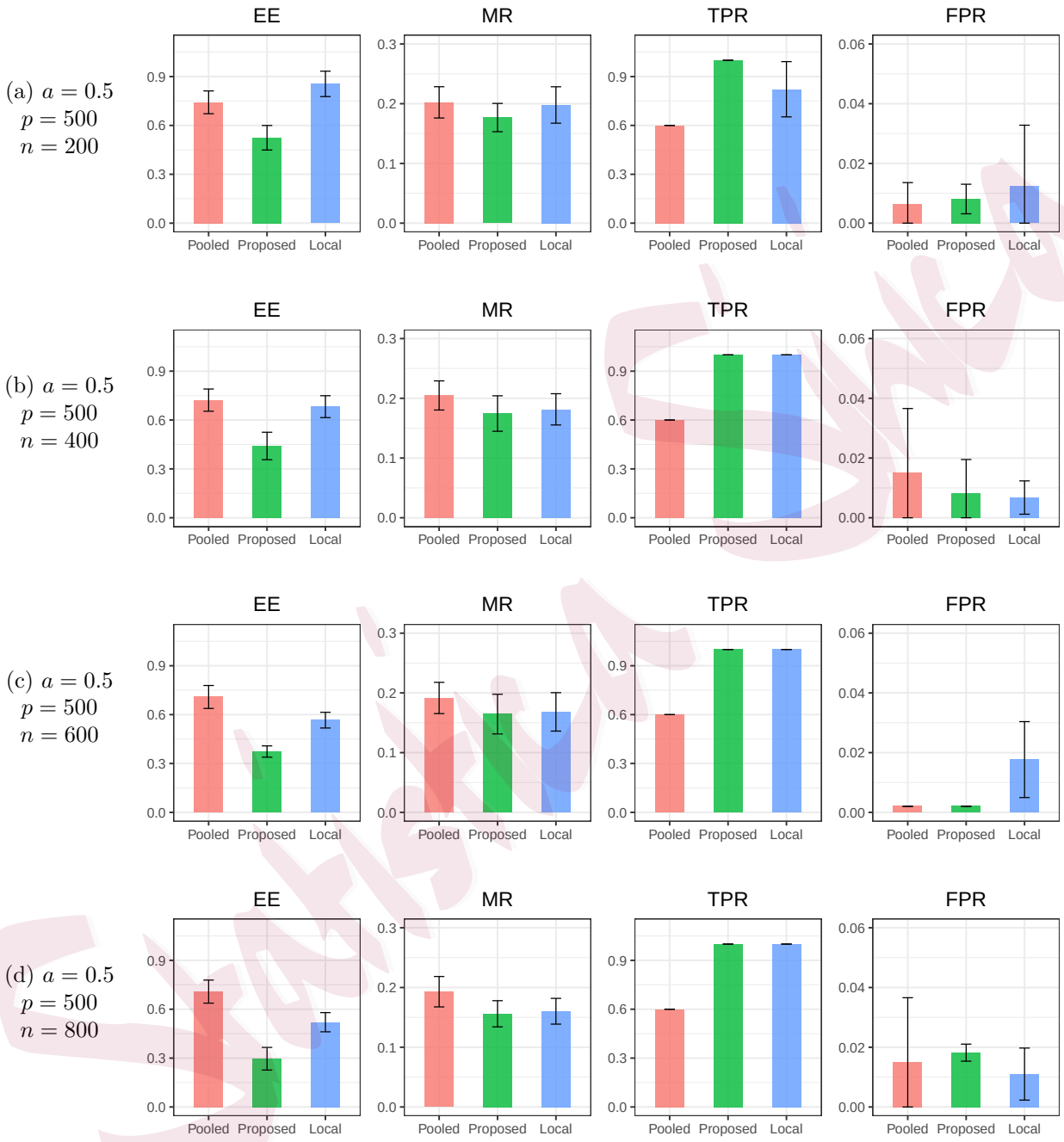


Figure 3: Comparison of different methods for varying $n \in \{200, 400, 600, 800\}$.

5. An Application to PISA Data

We apply the proposed multitask smoothed SVM to the Programme for International Student Assessment (PISA) 2012 dataset. PISA is a large-scale international educational assessment administered by the Organisation for Economic Co-operation and Development (OECD), designed to evaluate the reading, mathematics, and science skills of 15-year-old students. The original dataset contains 485,490 student records from 68 countries, with 635 predictors, including demographic attributes (e.g., country, gender, and parents' educational attainment), academic performance (e.g., mathematics scores), and attitudinal measures (e.g., attitudes toward mathematics). To reduce severe missingness and ill-conditioning, we restrict our analysis to countries with the largest sample sizes. Specifically, we select $M = 15$ countries with local sample size $n = 200$, and regard each country as a separate task and investigate whether collaborative margin-based classification can improve the prediction of low mathematics achievement while accommodating country-specific heterogeneity. The original response variable is the standardized mathematics score PV1MATH. We convert PV1MATH into a low-achievement indicator. Let $q_{0.25}$ be the lower quartile of PV1MATH computed from the pooled training sample. For student i in country m, we define

$$y_i^{(m)} = \begin{cases} 1, & \text{if } \mathrm{PV1MATH}_i^{(m)} \le q_{0.25}, \\ -1, & \text{otherwise.} \end{cases}$$

Thus, $y_i^{(m)} = 1$ corresponds to students at risk of low mathematics achievement. We compare three methods. The first method is the local estimator, which fits an independent sparse smoothed SVM for each country, ignoring information from other countries. The second

method is the pooled estimator, which fits a single sparse smoothed SVM after combining data from all countries. The third is the proposed multitask smoothed SVM. Since the true classification parameters are unknown in real data, we evaluate each method based on its prediction performance on a hold-out test set with size $n_{\text{test}} = 200$.

Table 1: Country-level misclassification rates on the PISA data. Results are averaged over 20 independent random splits, with standard errors in parentheses. Boxed entries indicate the lowest misclassification rate.

Country	Local	Pooled	Proposed
Mexico	0.24 (0.09)	0.23 (0.00)	0.16 (0.02)
Italy	0.09 (0.02)	0.06 (0.05)	0.08 (0.04)
Spain	0.15 (0.04)	0.15 (0.07)	0.13 (0.04)
Canada	0.05 (0.00)	0.01 (0.02)	0.04 (0.05)
Brazil	0.28 (0.11)	0.21 (0.09)	0.15 (0.11)
Australia	0.19 (0.12)	0.20 (0.00)	0.19 (0.05)
UK	0.05 (0.00)	0.06 (0.02)	0.06 (0.02)
UAE	0.33 (0.00)	0.21 (0.02)	0.23 (0.07)
Switzerland	0.05 (0.00)	0.09 (0.05)	0.06 (0.02)
Qatar	0.34 (0.09)	0.34 (0.05)	0.34 (0.02)
Colombia	0.34 (0.02)	0.26 (0.05)	0.29 (0.02)
Finland	0.10 (0.00)	0.08 (0.07)	0.06 (0.02)
Belgium	0.05 (0.00)	0.08 (0.04)	0.05 (0.04)
Denmark	0.09 (0.09)	0.10 (0.07)	0.11 (0.12)
Jordan	0.44 (0.12)	0.21 (0.02)	0.20 (0.04)

Table 1 shows that the proposed method achieves the lowest misclassification rate in eight of fifteen countries, including ties, with strict improvements over both alternatives in Mexico, Spain, Brazil, Finland, and Jordan. However, pooling performs better in Italy, Canada, UAE, and Colombia, while local fitting performs better in the UK, Switzerland, and Denmark. These differences are consistent with the tradeoff between information sharing and task-specific

flexibility: pooling can reduce variance when a country's classification pattern resembles the shared pattern, whereas local fitting can avoid borrowing bias when differences are substantial. The proposed method balances these effects, but regularization may shrink useful country-specific deviations. Neither joint estimation nor our theoretical guarantees imply uniformly better prediction in every country.

6. Conclusions

This paper develops a high-dimensional multitask smoothed SVM framework for heterogeneous binary classification. The central modeling device is a common/deviation decomposition of the task-specific linear classifiers. The common component captures features that are predictive across tasks, while the row-sparse deviation matrix captures features whose effects vary across tasks. This formulation provides a middle ground between fitting all tasks separately and imposing a single pooled classifier, and it is particularly suitable when interpretability and feature selection are important.

Methodologically, the estimator combines convolution smoothing of the hinge loss with a composite sparsity penalty. The smoothing step makes the empirical objective differentiable and enables Hessian-based local analysis, while the composite penalty induces sparsity in both the shared and task-specific parts. The resulting procedure remains faithful to the SVM margin-loss formulation but is more amenable to high-dimensional statistical analysis and gradient-based computation.

The theoretical results give estimation error bounds and support recovery guarantees for the proposed estimator. The error bound separates three sources of difficulty: the stochastic

fluctuation from high-dimensional empirical processes, the effective sparsity of the shared and heterogeneous active sets, and the approximation bias introduced by smoothing. The variable-selection result further shows that, under appropriate sparse eigenvalue, irrepresentability, signal-strength, and bandwidth conditions, the method can consistently recover both the shared active features and the features responsible for task heterogeneity. These results distinguish the present work from existing multitask SVM and kernel-learning methods, which are primarily algorithmic and typically do not provide high-dimensional coefficient-level theory for heterogeneous sparse linear SVMs.

Several directions remain open. First, the linear decision rule is useful for interpretation, but extensions to structured nonlinear feature maps or reproducing-kernel representations would broaden the scope of applications. Second, privacy-preserving or communication-efficient implementations are natural next steps for multisite studies where individual-level data cannot be pooled. Finally, it would be useful to develop post-selection inference for distinguishing shared predictive effects from task-specific deviations. Such developments would further strengthen the role of multitask smoothed SVMs as interpretable and theoretically grounded tools for high-dimensional heterogeneous classification.

Supplementary Materials

The Supplementary Material contains specific examples of smoothed hinge loss, detailed algorithmic descriptions, and the proofs for the main results.

Statistica Sinica: Newly accepted Paper

Acknowledgements

The authors thank the Editor, the Associate Editor, and two anonymous reviewers for their constructive comments and suggestions, which have greatly improved the paper. The research of Ling Peng is supported by Jiangxi Provincial Natural Science Foundation 20252BAC240162, and the China Postdoctoral Science Foundation 2024M751232. Yue Wang's research is supported by National Natural Science Foundation of China (Grant No.12501416), Anhui Provincial Natural Science Foundation (Grant No.2508085QA002), and the Fundamental Research Funds for the Central Universities (Grant No.WK2040250127). Xiaohui Liu's research is supported by the NSF of China (Grant No.12471257) and the NSF of the Science and Technology Department of Jiangxi Province (Grant No. 20243BCE51010), the NSSF of China (Grant No. 24&ZD069). This work is also supported by the Jiangxi Province Key Laboratory of Data Science in Finance and Economics (Grant No. 2024SSY03201).

References

Ando, R. K. and Zhang, T. "A framework for learning predictive structures from multiple tasks and unlabeled data." *Journal of Machine Learning Research*, 6(61):1817–1853 (2005).

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

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. "Simultaneous analysis of Lasso and Dantzig selector." *The Annals of Statistics*, 37(4):1705–1732 (2009).

Cai, T., Liu, M., and Xia, Y. "Individual data protected integrative regression analysis of high-dimensional heterogeneous data." *Journal of the American Statistical Association*, 117(540):2105–2119 (2022).

Chen, J., Liu, J., and Ye, J. "Learning incoherent sparse and low-rank patterns from multiple tasks." In *Proceedings*

Statistica Sinica: Newly accepted Paper

of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1179–1188. ACM (2010).

Chen, J., Zhou, J., and Ye, J. "Integrating low-rank and group-sparse structures for robust multi-task learning." In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 42–50. ACM (2011).

Cortes, C. and Vapnik, V. "Support-vector networks." *Machine Learning*, 20(3):273–297 (1995).

Cristianini, N. and Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge: Cambridge University Press (2000).

Duan, R., Ning, Y., and Chen, Y. "Heterogeneity-aware and communication-efficient distributed statistical inference." *Biometrika*, 109(1):67–83 (2022).

Evgeniou, T., Micchelli, C. A., and Pontil, M. "Learning multiple tasks with kernel methods." *Journal of Machine Learning Research*, 6(21):615–637 (2005).

Evgeniou, T. and Pontil, M. "Regularized multi-task learning." In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 109–117. ACM (2004).

Fernandes, M., Guerre, E., and Horta, E. "Smoothing quantile regressions." *Journal of Business & Economic Statistics*, 39(1):338–357 (2021).

Horowitz, J. L. "A smoothed maximum score estimator for the binary response model." *Econometrica*, 60(3):505–531 (1992).

Jacob, L., Vert, J.-P., and Bach, F. R. "Clustered multi-task learning: A convex formulation." In *Advances in Neural Information Processing Systems 21*, 745–752. Curran Associates, Inc. (2008).

Koo, J.-Y., Lee, Y., Kim, Y., and Park, C. "A Bahadur representation of the linear support vector machine." *Journal of Machine Learning Research*, 9(44):1343–1368 (2008).

Li, S., Cai, T. T., and Li, H. "Transfer learning for high-dimensional linear regression: Prediction, estimation and

Statistica Sinica: Newly accepted Paper

minimax optimality." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):149–173 (2022).

Lian, H. and Fan, Z. "Divide-and-conquer for debiased L_1-norm support vector machine in ultra-high dimensions." *Journal of Machine Learning Research*, 18(182):1–26 (2018).

Meier, L., van de Geer, S., and Bühlmann, P. "The group lasso for logistic regression." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71 (2008).

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers." *Statistical Science*, 27(4):538–557 (2012).

Obozinski, G., Wainwright, M. J., and Jordan, M. I. "Support union recovery in high-dimensional multivariate regression." *The Annals of Statistics*, 39(1):1–47 (2011).

Park, C., Kim, K.-R., Myung, R., and Koo, J.-Y. "Oracle properties of SCAD-penalized support vector machine." *Journal of Statistical Planning and Inference*, 142(8):2257–2270 (2012).

Peng, B., Wang, L., and Wu, Y. "An error bound for L_1-norm support vector machine coefficients in ultra-high dimension." *Journal of Machine Learning Research*, 17(233):1–26 (2016).

Rybak, J., Battey, H., and Zhou, W.-X. "On inference for the support vector machine." *Journal of Machine Learning Research*, 26(85):1–54 (2025).

Schölkopf, B. and Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press (2002).

Steinwart, I. and Christmann, A. *Support Vector Machines*. Information Science and Statistics. New York: Springer (2008).

Tan, K. M., Wang, L., and Zhou, W.-X. "High-dimensional quantile regression: Convolution smoothing and concave regularization." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):205–233 (2022).

Statistica Sinica: Newly accepted Paper

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. "Sparsity and smoothness via the fused lasso." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108 (2005).

Vapnik, V. N. *The Nature of Statistical Learning Theory*. Information Science and Statistics. New York: Springer, 2 edition (2000).

Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press (2019).

Wang, B., Zhou, L., Gu, Y., and Zou, H. "Density-convoluted support vector machines for high-dimensional classification." *IEEE Transactions on Information Theory*, 69(4):2523–2536 (2023).

Wang, B., Zhou, L., Yang, J., and Mai, Q. "Density-convoluted tensor support vector machines." *Statistics and Its Interface*, 17(2):231–247 (2024).

Wang, L., Zhu, J., and Zou, H. "The doubly regularized support vector machine." *Statistica Sinica*, 16(2):589–615 (2006).

Wang, Y. and Lian, H. "Sparse additive support vector machines in bounded variation space." *Information and Inference: A Journal of the IMA*, 13(1):iaae003 (2024).

Wang, Y., Lian, H., and Zhou, W. "Gbs-SVM: A linearly convergent gossip algorithm for high-dimensional classification." *Available at SSRN 6270798* (2026).

Wolfson, M., Wallace, S. E., Masca, N., Rowe, G., Sheehan, N. A., Ferretti, V., LaFlamme, P., Tobin, M. D., Macleod, J., Little, J., Fortier, I., Knoppers, B. M., and Burton, P. R. "DataSHIELD: Resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data." *International Journal of Epidemiology*, 39(5):1372–1382 (2010).

Yuan, M. and Lin, Y. "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67 (2006).

Zhang, Y. and Yang, Q. "A survey on multi-task learning." *IEEE Transactions on Knowledge and Data Engineering*,

Statistica Sinica: Newly accepted Paper

34(12):5586–5609 (2022).

Ling Peng

School of Statistics and Data Science and Philosophy and Social Sciences Laboratory of Data Science in Finance and Economics at the Ministry of Education, Jiangxi University of Finance and Economics, Nanchang, China

E-mail: pengling@jxufe.edu.cn

Yue Wang

Faculty of Business for Science and Technology, School of Management, USTC, Hefei, China

E-mail: wangyue5@ustc.edu.cn

Xiaohui Liu

Philosophy and Social Sciences Laboratory of Data Science in Finance and Economics at the Ministry of Education and School of Statistics and Data Science Jiangxi University of Finance and Economics, Nanchang, China

E-mail: liuxiaohui@jxufe.edu.cn

Heng Lian

City University of Hong Kong Shenzhen Research Institute, Shenzhen, China

Department of Mathematics, City University of Hong Kong, Hong Kong, China

E-mail: henglian@cityu.edu.hk