Statistics and Probability Letters 240 (2027) 110933

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Generalized linear multiple kernel learning for high-dimensional image data☆

Qiyuan Shi, Jian Kang, Yi Li *

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

ARTICLE INFO

Keywords:
Chest X-ray
Feature extraction
Generalized linear models
Multiple kernel learning

ABSTRACT

We propose GLIMARK, a sparse generalized linear multiple-kernel framework for high-dimensional structured predictors. It extends additive kernel learning to exponential-family outcomes, supports multi-view and hierarchical feature representations, improves scalability via Nyström approximation, and yields interpretable feature-kernel and representer selection.

1. Introduction

High-dimensional data in image analysis often arise from heterogeneous sources, such as segmentations or anatomical regions. Image features may be extracted using radiomic (Van Griethuysen et al., 2017; Sun et al., 2023) or deep learning methods (Tanida et al., 2023). These strategies often share two characteristics. First, the resulting predictors are often high-dimensional. Second, they are often structurally heterogeneous, with features organized by regions, filters, feature classes, or other meaningful partitions. These characteristics create two related challenges: capturing complex nonlinear relationships and accommodating heterogeneous feature representations across partitions. Kernel methods provide a natural framework for the first challenge, since they model nonlinear relationships through similarities between observations without requiring an explicit finite-dimensional transformation. For the second challenge, when predictors are partitioned into heterogeneous components, a single kernel may be too restrictive.

Multiple Kernel Learning (MKL) (Gönen and Alpaydın, 2011) addresses this by combining pre-defined kernels to construct a more flexible representation. This construction is particularly useful when predictors are partitioned into components that capture distinct aspects of the data. By assigning kernels across partitions and learning their weights from data, one can obtain a flexible and data-driven representation; when the number of partitions is large, a common set of kernels may be used across all partitions, which also connects to multi-view learning (Yan et al., 2021). Inspired by this paradigm, we refer to a kernel-partition combination as a "view" (Fig. 1(a)). The usefulness of kernel methods is closely tied to the representer theorem (Wahba and Wang, 2019), which states that for many regularized learning problems in a reproducing kernel Hilbert space (RKHS), the solution can be expressed as a finite linear combination of kernel evaluations at the training points. Such representations are widely used in kernel ridge regression (Lin et al., 2020), kernel Cox models (Rong et al., 2024), and deep learning (Unser, 2019). This makes kernel methods practical in high-dimensional settings and provides a natural interpretation in terms of weighted similarities to training observations.

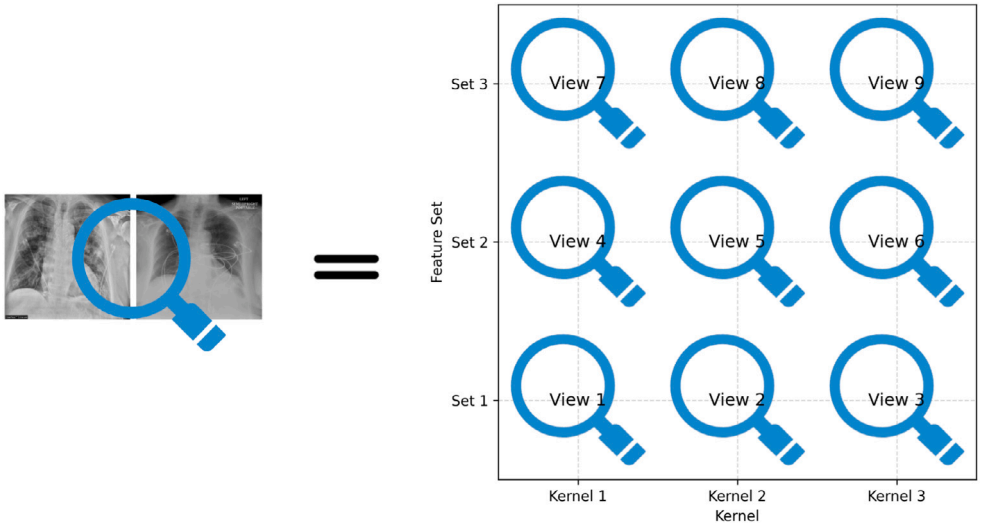
However, most MKL extensions in dimension-reduction approaches such as MKL-DR (Lin et al., 2010) often rely on projection or embedding, which can compromise interpretability (Briscik et al., 2023). To address this issue, we focus on Multiple Additive Regression Kernels (MARK) (Bennett et al., 2002), which is based on column generation (Demiriz et al., 2002). Instead of searching for an optimal kernel, MARK selects feature-kernel combinations directly, thereby preserving the interpretability of the original

☆ This article is part of a Special issue entitled: 'STAPRO_High-Dimensional Data Analytics' published in Statistics and Probability Letters.

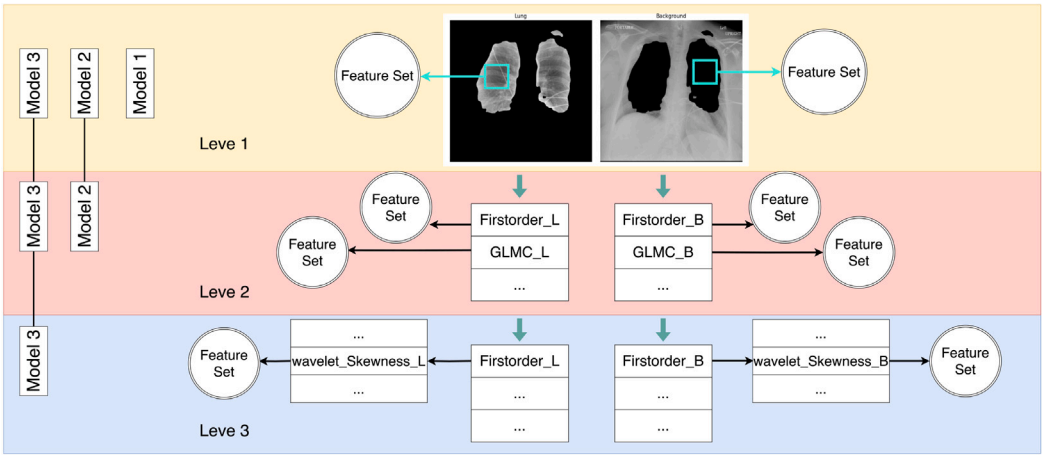
* Corresponding author.
E-mail address: yili@umich.edu (Y. Li).

https://doi.org/10.1016/j.spl.2026.110933

Available online 19 August 2026



(a) A multi-view comparison of two images.



(b) Hierarchical model structures.

Fig. 1. Conceptual overview of the proposed framework. Panel (a) illustrates multi-view kernel construction across feature sources, and panel (b) illustrates the hierarchical image-based refinement used in the CXR application.

features. Nonetheless, MARK still has important limitations. Because it relies on squared loss, its applicability to binary and other non-Gaussian outcomes is unclear. In addition, although the method scales nearly linearly with sample size, it becomes burdensome when the number of feature-kernel combinations is large. This issue is especially important in structured high-dimensional settings, where predictors may be partitioned into many meaningful groups and different groups may require distinct kernel representations. Given a sample size of N, P partitions of features, and Q types of kernels, the full kernel matrix $\boldsymbol{K}$ consists of NPQ columns, which can create substantial computational burden.

To address these limitations, we propose *Generalized Linear Models with Integrated Multiple Additive Regression with Kernels* (GLIMARK). GLIMARK introduces a likelihood-based loss function within the generalized linear model framework, accommodating a range of outcomes including binary, Poisson, and Gaussian. We further propose a hierarchical framework for images with multiple sources of information, allowing the method to explore feature patterns at multiple scales while balancing model complexity and pattern learning capacity. To improve scalability when the number of candidate feature-kernel combinations is large, we can optionally incorporate a Nyström approximation into the algorithm. In addition, the modification with the $\mathcal{H}_k$-norm provides a more rigorous explanation aligned with the representer theorem. The resulting framework preserves the sparse and interpretable selection mechanism of MARK while extending it to heterogeneous high-dimensional data and non-Gaussian outcomes.

The remainder of the paper is organized as follows. Section 2 introduces the proposed GLIMARK. Section 3 presents simulation studies. Section 4 applies the method to real image data. Section 5 concludes with discussion.

2. Generalized linear models with integrated multiple additive regression with kernels (GLIMARK)

2.1. Method

To accommodate non-Gaussian outcome data, we extend additive kernel learning to the generalized linear model (GLM) setting (Bennett et al., 2002; Zhu and Hastie, 2005). Assume y_i follows an exponential-family distribution with density

$$p(y|\theta) = \exp\left(y\theta - b_{glm}(\theta) + c_{glm}(y)\right), \tag{1}$$

where we omit the dispersion parameter as in Fei and Li (2021). Under the canonical link, let

$$\theta_i = f_i = g_{glm}(\mu_i), \qquad \mu_i = \mathbb{E}[Y_i \mid \boldsymbol{z}_i].$$

We estimate f by minimizing the regularized empirical risk

$$H = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i f_i - b_{glm}(f_i)\right] + \lambda P(\|f\|_{\mathcal{H}_k}), \tag{2}$$

where $\lambda > 0$ is a regularization parameter and $\|f\|_{\mathcal{H}_k}$ is the RKHS norm.

With P feature partitions and Q kernels, a standard MKL formulation is

$$f(\boldsymbol{z}) = \sum_{j=1}^{N}\alpha_j\sum_{q=1}^{Q}\sum_{p=1}^{P}\mu_{p,q}k_q(\boldsymbol{z}_j^p, \boldsymbol{z}^p) + b.$$

However, estimating both α_j and $\mu_{p,q}$ requires additional constraints and can be computationally demanding. Following MARK (Bennett et al., 2002), we instead use the additive representation

$$f(\boldsymbol{z}) = \sum_{j=1}^{N}\sum_{q=1}^{Q}\sum_{p=1}^{P}\alpha_{j,p,q}k_q(\boldsymbol{z}_j^p, \boldsymbol{z}^p) + b, \tag{3}$$

which directly selects patient-view combinations.

Let $\boldsymbol{K}$ denote the concatenated design matrix of kernel blocks across all partitions and kernels:

$$\boldsymbol{K} = \begin{bmatrix}\boldsymbol{K}_{1,1} & \cdots & \boldsymbol{K}_{1,Q} & \boldsymbol{K}_{2,1} & \cdots & \boldsymbol{K}_{2,Q} & \cdots & \boldsymbol{K}_{P,1} & \cdots & \boldsymbol{K}_{P,Q}\end{bmatrix}, \tag{4}$$

where each block $\boldsymbol{K}_{p,q}$ is an $N \times N$ matrix with entries $k_q(\boldsymbol{z}_i^p, \boldsymbol{z}_j^p)$. Thus, each column of $K_{p,q}$ corresponds to one patient under the view defined by partition p and kernel q. Writing $J = NPQ$, model (3) becomes

$$f_i = f(\boldsymbol{z}_i) = \sum_{j=1}^{J}\alpha_j[\boldsymbol{K}]_{i,j} + b. \tag{5}$$

GLIMARK builds a sparse model by forward column generation. Starting from $\boldsymbol{\alpha}^{(0)} = \boldsymbol{0}$, at each iteration it selects one column of $\boldsymbol{K}$ and then updates the selected coefficients to reduce (2). Selection is based on the magnitude of the partial derivative

$$\begin{aligned}\frac{\partial H(\boldsymbol{\alpha}, b)}{\partial \alpha_j} = &-\frac{1}{N}\sum_{i=1}^{N}\left[y_i - \frac{\partial b_{glm}(f_i(\boldsymbol{\alpha}, b))}{\partial f_i}\right][\boldsymbol{K}]_{i,j} \\ &+ \lambda\frac{\partial P(\|f(\boldsymbol{\alpha}, b)\|_{\mathcal{H}_k})}{\partial \alpha_j},\end{aligned} \tag{6}$$

evaluated at the current iterate $(\boldsymbol{\alpha}^{(t)}, b^{(t)})$. Let $S^{(t)}$ denote the selected index set, with $S^{(0)} = \emptyset$. At step $t+1$, we select

$$j^{*(t+1)} = \arg\max_{j \notin S^{(t)}}\left|\frac{\partial H(\boldsymbol{\alpha}, b)}{\partial \alpha_j}\right|, \tag{7}$$

and update the model using

$$f_i = \sum_{j \in S^{(t+1)}}\alpha_j[\boldsymbol{K}]_{i,j} + b. \tag{8}$$

The coefficient updates are obtained by gradient-based optimization such as Adam (Kingma, 2014). To improve scalability when NPQ is large, we optionally incorporate a Nyström approximation (Drineas et al., 2005), replacing each $N \times N$ kernel block by an $N \times d$ landmark-based representation and reducing storage and kernel construction from $\mathcal{O}(N^2PQ)$ to $\mathcal{O}(dNPQ)$. Because the approximation is built from observed-sample columns, it remains well aligned with representer-based interpretation. The full algorithm is summarized in Appendix A.

For example, when $P(\cdot)$ is the square function, the objective becomes

$$H(\boldsymbol{\alpha}, b) = \begin{cases}-\frac{1}{N}\sum_{i=1}^{N}\left[y_i f_i - \log(1+\exp(f_i))\right] + \lambda\|f\|_{\mathcal{H}_k}^2, & \text{(Binomial)},\\ -\frac{1}{N}\sum_{i=1}^{N}\left[y_i f_i - \exp(f_i)\right] + \lambda\|f\|_{\mathcal{H}_k}^2, & \text{(Poisson)}.\end{cases}$$

It should be noted that the representer theorem provides the theoretical foundation for reducing the infinite-dimensional function-space problem to a finite-dimensional representation, while GLIMARK constructs a parsimonious approximation to the full penalized solution through iterative column generation.

3. Simulation study

We conducted seven simulation scenarios, with 100 replicates per scenario, to evaluate two aspects of GLIMARK and to compare its performance with commonly used benchmark models, including ℓ_1- and ℓ_2-penalized logistic regression, random forests, neural networks, XGBoost, and LightGBM.

The first aim was to evaluate recovery under linear, nonlinear, and interaction-generating mechanisms. We found GLIMARK provided clear recovery of nonlinear and interaction structures when suitable kernels were specified, while also showing sensitivity to kernel misspecification. The second aim was to examine predictive-performance patterns across the two-dimensional grid of λ and T, as well as to compare GLIMARK with the benchmark methods. GLIMARK achieved competitive performance for the binary outcome, while XGBoost and LightGBM attained the highest AUCs and generally stronger classification metrics. For the count outcome, GLIMARK achieved the lowest RMSE among all evaluated methods. Full details are provided in Appendix B.

4. Application to precision health COVID-19 CXR data

We applied GLIMARK to COVID-19 chest X-ray data from the Precision Health database, including 2396 patients and 2046 radiomic features, with 756 patients experiencing ICU escalation within 30 days of initial diagnosis. Cohort construction, preprocessing, radiomic feature extraction, and descriptive summaries are provided in Appendix C.

The data were divided into an 80% development set and a 20% holdout test set. Within the development set, we used 100 bootstrap resamples and a staged hierarchical refinement from segmentation-level information to feature classes and individual features to identify three models, M1–M3. These models and the benchmark methods were evaluated by 5-fold cross-validation in the development set, and final performance was assessed on the holdout set. The primary analysis used the full kernel, with model development, performance evaluation, and Nyström sensitivity analyses reported in Appendices C.2–C.4, respectively. On the 20% holdout set, M3 achieved the highest AUC among all evaluated models (0.835) and was considered the primary model.

In terms of interpretation, the M3 GLIMARK concentrated signal on a small number of feature classes, with First-Order, NGTDM, and Shape2D emerging as dominant contributors. The prominence of First-Order and Shape2D is consistent with prior radiology findings (Varghese et al., 2021; Al-Zyoud et al., 2023; Hiremath et al., 2024), while NGTDM was additionally highlighted by our analysis. The apparent contribution of the non-lung region may reflect clinically relevant extrapulmonary information but may also capture medical devices, positioning, cropping, or other acquisition-related artifacts and should therefore be interpreted cautiously. GLIMARK also identifies influential representer patients, providing an example-based interpretation of the fitted model. Detailed model comparisons, feature-importance plots, and representer visualizations are deferred to Appendix C.

5. Discussion

GLIMARK is flexible in exploring feature structure across multiple scales through hierarchical column generation. Higher-level feature groups can capture broader interaction patterns, whereas lower-level expansions enable more targeted feature selection. Unlike projection-based approaches such as KPCA, GLIMARK with Nyström approximation does not remap predictors before modeling, which helps retain interpretability when handcrafted features are used.

Several directions remain for future work. Methodologically, formal convergence guarantees may be established, and GLIMARK may be extended to noncanonical links and overdispersed outcomes. Application-wise, although our CXR application uses a multi-view strategy, it does not fully exploit the potential of GLIMARK for richer heterogeneous structures. More localized region definitions, potentially informed by transformer-based attention maps (Dosovitskiy et al., 2020), may support a more systematic multi-view design across image perspectives, regions of interest, filters, and feature classes. In addition, the current kernel construction is not exhaustive. Domain-specific kernels have been extensively studied in other settings (Shervashidze et al., 2011; Lodhi et al., 2002), but analogous kernels tailored to CXR data remain limited. Future work may integrate radiomic and deep-learning features within GLIMARK and evaluate generalizability beyond a single health system.

Funding

This work was supported by the National Cancer Institute under grant numbers R01CA269398 and R01CA249096.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spl.2026.110933.

Data availability

Restrictions apply to the availability of the data, which were used under license for this study. Data are available upon request from Michigan DataDirect (PHDataHelp@umich.edu) with the necessary permissions.

References

Al-Zyoud, W., Erekat, D., Saraiji, R., 2023. COVID-19 chest X-ray image analysis by threshold-based segmentation. Heliyon 9 (3).

Bennett, K.P., Momma, M., Embrechts, M.J., 2002. MARK: A boosting algorithm for heterogeneous kernel models. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 24–31.

Briscik, M., Dillies, M.-A., Déjean, S., 2023. Improvement of variables interpretability in kernel PCA. BMC Bioinformatics 24 (1), 282.

Demiriz, A., Bennett, K.P., Shawe-Taylor, J., 2002. Linear programming boosting via column generation. Mach. Learn. 46, 225–254.

Dosovitskiy, A., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Drineas, P., Mahoney, M.W., Cristianini, N., 2005. On the Nyström method for approximating a gram matrix for improved kernel-based learning.. J. Mach. Learn. Res. 6 (12).

Fei, Z., Li, Y., 2021. Estimation and inference for high dimensional generalized linear models: A splitting and smoothing approach. J. Mach. Learn. Res. 22 (58), 1–32.

Gönen, M., Alpaydın, E., 2011. Multiple kernel learning algorithms. J. Mach. Learn. Res. 12, 2211–2268.

Hiremath, A., Viswanathan, V.S., Bera, K., Shiradkar, R., Yuan, L., Armitage, K., Gilkeson, R., Ji, M., Fu, P., Gupta, A., Lu, C., Madabhushi, A., 2024. Deep learning reveals lung shape differences on baseline chest CT between mild and severe COVID-19: A multi-site retrospective study. Comput. Biol. Med. 177, 108643. http://dx.doi.org/10.1016/j.compbiomed.2024.108643.

Kingma, D.P., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Lin, Y.-Y., Liu, T.-L., Fuh, C.-S., 2010. Multiple kernel learning for dimensionality reduction. IEEE Trans. Pattern Anal. Mach. Intell. 33 (6), 1147–1160.

Lin, S.-B., Wang, D., Zhou, D.-X., 2020. Distributed kernel ridge regression with communications. J. Mach. Learn. Res. 21 (93), 1–38.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., Watkins, C., 2002. Text classification using string kernels. J. Mach. Learn. Res. 2 (Feb), 419–444.

Rong, Y., Zhao, S.D., Zheng, X., Li, Y., 2024. Kernel Cox partially linear regression: Building predictive models for cancer patients' survival. Stat. Med. 43 (1), 1–15.

Shervashidze, N., Schweitzer, P., Van Leeuwen, E.J., Mehlhorn, K., Borgwardt, K.M., 2011. Weisfeiler-lehman graph kernels.. J. Mach. Learn. Res. 12 (9).

Sun, Y., et al., 2023. Use of machine learning to assess the prognostic utility of radiomic features for in-hospital COVID-19 mortality. Sci. Rep. 13 (1), 7318.

Tanida, T., Müller, P., Kaissis, G., Rueckert, D., 2023. Interactive and explainable region-guided radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7433–7442.

Unser, M., 2019. A representer theorem for deep neural networks. J. Mach. Learn. Res. 20 (110), 1–30.

Van Griethuysen, J.J., et al., 2017. Computational radiomics system to decode the radiographic phenotype. Cancer Res. 77 (21), e104–e107.

Varghese, B.A., et al., 2021. Predicting clinical outcomes in COVID-19 using radiomics on chest radiographs. Br. J. Radiol. 94 (1126), 20210221.

Wahba, G., Wang, Y., 2019. Representer theorem. Wiley StatsRef: Stat. Ref. Online 1–11.

Yan, X., Hu, S., Mao, Y., Ye, Y., Yu, H., 2021. Deep multi-view learning methods: A review. Neurocomputing 448, 106–129.

Zhu, J., Hastie, T., 2005. Kernel logistic regression and the import vector machine. J. Comput. Graph. Statist. 14 (1), 185–205.