IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 22, Issue 3, Ser. 1 (May. – June. 2026), PP 15-25
www.iosrjournals.org

A Statistical Review Of Methods For Addressing The P>N Problem: Logistic Regression As A Foundational Classification Framework (2015–2025)

Nada Hussein Tali, Hanan Abbas Hamza
(Department Of Statics, College Of Administration & Economic / University Of Sumer, Iraq)
(Department Of Studies And Planning/ University Of Babylon, Iraq)

Abstract:
Logistic regression (LR) is a popular machine learning (ML) technique for predicting binary outcomes. It has numerous applications in fields from healthcare to customer churn. However, the performance of LR can deteriorate when the number of features (p) exceeds the number of samples (n) a problem frequently encountered in high-dimensional datasets. Under such circumstances, overfitting occurs which degrades the performance of LR in future predictions. This issue has been extensively studied in the ML and statistics research communities. Numerous regularization techniques have been proposed to improve LR in high-dimensional settings. The fundamental idea of all methods is to penalize certain features with regularization terms to obtain a more parsimonious model. Hence, the primary motivation for these techniques is to eliminate irrelevant features for performance improvement. Regularized LRs can be categorized as: (1) Lasso (L1), which tends to perform feature selection due to the sparsity constraint on the coefficients in the model. (2) Ridge (L2), which shrinks the coefficients of all selected features proportionally to their importance. (3) Elastic Net, which mixes the penalties imposed by Lasso and Ridge regularization for model sparsity and feature selection, respectively. Lasso provides a starting point for selecting features. In general, Ridge proves to be a better algorithm for the modeling accuracy of the retained features, making it a good alternative when the dimensionality is high, and the differences in the importance of the features are very small. However, a combination of the two techniques (Lasso and Ridge) yields a regularized LR that exhibits the advantages of both techniques: feature selection and model sparsity, resulting in a model that is computationally efficient, accurately calibrated, and has less uncertainty compared to traditional LR. Implementation of regularized LR can be computationally intensive due to the need for hyperparameter tuning and model fitting. However, the techniques provide greater model performance than traditional ML algorithms such as support vector machines (SVMs) and Random Forests.

Key Word: Logistic regression *;* Ridge*;* Lasso*;* Elastic Net*.*

Date of Submission: 26-04-2026 Date of Acceptance: 06-05-2026

I. Introduction

In supervised learning, a predictive task consists of inferring the values of a response variable based on the values of one or more covariates. Paired observations of response and covariate values constitute a dataset, which is a finite and fixed realization of an underlying data-generating process. Regression is a predictive task where the response variable is continuous, while classification refers to a predictive task where the response variable is categorical. Many real-world classification datasets exhibit high dimensionality, where the number of covariates p exceeds the number of paired observations n ($p > n$). Classification under such conditions is often referred to as the $p > n$ problem. Logistic regression is a popular modelling and inference technique that seeks to describe the relationship between a categorical response variable and one or more covariate variables from a statistical point of view (Sur & J. Candes, 2018). However, both high-dimensional and low-dimensional datasets are commonly encountered in practice, so it is of interest to develop methods for assessing the performance of a particular classification method in a truly high-dimensional setting, when $p > n$.

II. Background And Problem Formulation

The dimension of covariates frequently exceeds sample size in modern datasets, leading to the p>n problem. Formally, let y be a n-dimensional binary response vector, X a $n \times p$ covariate matrix where p exceeds n, and g its link function. The dependent structure of the outcome can differ across disciplines. A popular approach within the logistic regression framework assumes a linear model based on sigmoidal functions (Sur et al., 2017). The model is not without debate; alternative methods allow certain assumptions to be relaxed while

maintaining classification with good performance. Logistic regression remains a useful starting point when addressing high-dimensional problems.

Evaluating a given classifier in high dimensionality poses challenges. Independent datasets containing observations under the same domain often do not exist, complicating the assessment of a method's generalizability, stability, or interpretability. Conclusions drawn from a dataset can be confounded by multiple factors. Careful choice of evaluation criteria, data generation and pre-processing procedures, and reporting of results increase the likelihood of unbiased, interpretable conclusions.

Foundations of Logistic Regression in High-Dimensional Settings

Classification elucidates the association between a response variable and one or more predictors, facilitating predictions for new observations (Salehi et al., 2019). The logistic and linear regression models stand as foundational works within this expansive field, receiving widespread across disciplines. Establishing a suitable baseline framework for high-dimensional settings, particularly prevalent in domains ranging from genomics to imaging, remains a notable challenge (Sur et al., 2017). The $p > n$ regime represents a quintessential high-dimensional setting, wherein the number of covariates exceeds the number of observations questioning the feasible identification of the true underlying signal.

Logistic regression emerges as a promising contender for a baseline $p > n$ classification framework. The need for model selection, during which available covariates are restricted to a subset prior to invoking the classifier, creates a critical dilemma prior to model fitting the framework of choice. In contrast to ensemble techniques like Random Forests, which necessitate entire training sets, logistic-type classifiers possess a distinctive advantage: computational speed sufficient model fits within the constraints specified by the selected $p > n$ algorithm.

Regularization Techniques for p>n Classification

Recent studies indicate that the high-dimensional approach to logistic-regression classification within the statistical literature from 2015 to 2020 remains relevant: the substantial growth of substantial datasets necessitates methods capable of accommodating dimensions exceeding the sample size. Within that corpus, a shared modelling perspective motivates the selection of logistic regression as a baseline classifier. High-dimensional logistic regression possesses distinctive foundational aspects which retain practical significance and have elicited diverse regularization attempts. By considering logistic regression within the context of high-dimensionality efforts, a systematic overview of the salient methods addressing regularization for p>n classification is presented (Salehi et al., 2019).

L1 and L2 Regularization

Correct penalization of high-dimensional estimation problems must conform to data sparsity patterns since greater noise magnifies less informative features. Sparse, multi-frequent, and hierarchical signals are ubiquitous across research domains. In high-dimensional settings, readily accessible, modifiable, and well-studied alternatives support temporal features through widely available software. Existing methodologies begin by fixing covariates prior to formulating prediction models that omit interactions and non-linear transformations, leading to unobserved or unaddressed feature correlations, motivation for ordinary sequential designs, and widespread adoption of the baseline model. The interplay between features guides choice of classical, likelihood-based, underpinning models, and procedure characteristics provide a foundation for delineating procedures. The consistent probability of using incorrect models, including those based on features chosen ex-ante, merely correlates with state-of-the-art gauge solutions. Considering the nature of commonly encountered datasets favours probit siblings and auxiliary developments specific to the dataset. Concerningly, conditional, knot-variate choices depart from the generating process in high-dimensional datasets.

Modern procedures fail either to maintain standard nonlinear, broadening perspectives on readily available baseline models or to permit judicious choices among freely available, modifiable, established procedures equipped with stepwise temporal correlations. A coherent view across disciplines thus evolves towards high-dimensional datasets of dimension p exceeding size n. Conditional and transformation-specific models operating on principal and independent components seldom conform to widely available datasets. Kernel-density capacity and variation-based policies continue to direct attention towards logistic regression under widely available software and well-researched properties.

Extensive sampling reveals focal attention on the intermediary landscape preambles accessible via ordinary stepwise schemes. Variants striving to fulfil, cumulatively provide initial expressions consistent with widespread sampling prior to shifting focus to elaborate, generative auxiliary datasets operating in the joint dataset such that a p>n emerging temporal state assists in transitioning among diverse disciplines (An & Zhang, 2020) ; (Salehi et al., 2019) ; (Liu, 2014).

Conditional probabilities and directly related probabilities under a proper prior formulate an equally viable distanced policy. Decenário, posterior sampling reduces the characteristic of the expected outline to a deterministic entity tracing the p>n landscape from contours already widely sampled. The independent mode assumes that features p align with the data-generating architecture and reduces sophisticated modelling to the exploration of few selected aspects. Linear generalizations systematically pose a similar expansion challenge.

Elastic Net and Hybrid Penalties

Penalization removes the absolute value of the coefficients from the objective function in logistic regression, leading to a non-convex optimization problem. The elastic net method allows for grouped variable selection through a hybrid penalization approach, which combines both L1 and L2 penalties. Grouped variable selection is highly desired in general, but particularly for isotonic regression, where the variables of interest may arise in groups (Sevinc Kurnaz et al., 2017).

The elastic net penalty introduces a trade-off between a grouped and an ungrouped variable selection; the elastic net still works well even with unrelated features because it is capable of restoring a good level of fitting without setting coefficients to be large (Balcan et al., 2022). Even though hyper-parameter tuning is necessary for logistic regression with elastic net penalties, a solution still exhibits semi-elastic net solutions across multiple dataset instances, enabling the sensitivity of hyper-parameter tuning to be reduced. The elastic net penalty can thus be tuned without requiring extensive grid-search tuning. SpeeDee, a multi-task learning approach for dynamic regularization in supervised learning, can employ a flexible selection of the elastic net hyper-parameters.

Sparse Bayesian Methods

Consider the conditional density function of a response variable $y \in Y$ given an observed input $x \in X$. The likelihood function takes the form of $p(y|x)$, where an observation consists of a tuple (y, x) of variables constrained by specific probability laws. When standard data-producing mechanisms yield observations fitting such a model, it is referred to as a generative environment, enabling a viable probabilistic model to infer from partial observations. Otherwise, with data generated by a different process, it constitutes a non-generative environment, and a viable probabilistic model cannot be constructed.

A necessary condition for observing a response y is the presence of an actual variable from model neighborhoods; stable redistributions cannot observe fresh neighborhoods without prior sightings. A standard approach to mitigate overfitting risk is the sparse model, targeting $P(y|x)$ estimation without the necessity of observing x. Bayesian analysis interprets the posterior distribution obtained from observations as a quantifications of initial degrees of belief within a predefined hypothesis class. When $N = 0$, it concerns the prior distribution, indicating what is known about the model before seeing any fresh data. The posterior serves as a summarization of model uncertainty induced by population data and may effectively constrains the search of consistent estimation from a candidate model; ideally, the family of distributions continues to provide rich modeling options even when T is increasing and observed y have remained unchanged. When $T > 0$ is observed, the approach inspires prior determination on candidate parametrization or model. (Chakraborty, 2019).

Stability and Model Selection Criteria

Stability selection is a two-step variable selection approach, which incorporates the stability of selected variables into the selection criterion. Stability selection is based on a subsampling scheme applied to the original dataset, where subsets of data are used to train the model in order to assess the importance of a variable (Fang et al., 2013). The criterion for variable importance can be based on either the area under the receiver operating characteristic (ROC) curve (AUC) or the prediction error rate. The results are combined across different splits of the data to obtain the final importance measure.

Several further extensions and variants of criteria of stability selection are summarized and implementation details are discussed. Information criteria have been modified to account for the situation $p > n$ (Pokarowski et al., 2012). The goal is to select a model f among the collection of possible models $f1, \ldots, fm \in F$ that best fits the observation D. The model complexity is included in the model selection condition through Kullback-Leibler (K-L) type penalisation. For all $1 < q < \infty$, a model selection criterion is proposed based on the observation D, and satisfying certain properties. Given a collection of candidate models robust model selection based on K-L type minimum description length principle is also investigated; the asymptotic property of the integrated noise-and-parameter model selection criterion is established.

III. Dimensionality Reduction Approaches

Addressing the curse of dimensionality in regression analysis demands effective dimensionality-reduction strategies that mitigate the p>n challenge without sacrificing essential information. Nevertheless, conventional feature-ranking and filtering methods exhibit limitations that necessitate recourse to other

screening instruments, such as screening thresholds, sure independence screening, or error-rate control (Liu, 2015). The high cost of directly estimating high-dimensional covariance matrices underp and the ubiquity of uncorrelated principal components in high-dimensional text data propel the use of linear approximation via principal component analysis (Weng & S. Young, 2017). This technique enables projection onto low-dimensional linear subspaces that capture considerable variation. Geometric alterations to principal-component-regression residuals pursue principal directions preserving maximum predictive information. Rotation-invariant projections facilitate sparse vector recovery along generally unknown geometric directions. Partial least squares further relaxes the premise of maximum variation as a sufficient condition for predictive information.

In conjunction with linear dimension-reduction techniques, independent-component-analysis algorithms translate high-dimensional observations into independent sources. Model identifiability and interpretation under stringent independence and invertibility stipulations prove essential for establishing theoretical underpinnings, although these restrictions vary according to prior assumptions governing data generation. Nonlinear transformation into low-dimensional spaces, potentially via deep neural networks, constitutes another dimension-reduction strategy that remains amenable to logistic regression integration.

Feature Screening

Feature screening constitutes the foundation of many variable-selection techniques (Roy et al., 2022) and reduces the search space for more complex models, thus facilitating the application of high-dimensional generalizations of classical statistical classifiers (Zhang et al., 2020). A screening threshold is a decision rule that separates important from unimportant variables; features passing the threshold are retained for subsequent analysis, while those falling below are discarded. Forgetting any variable whose effect is larger than some fixed threshold yields a method known as sure independent screening (SIS). Another common thresholding approach is to rank variables according to some relevant criterion, where the cleared dimensions are those corresponding to the k highest, or largest, scores (Liu et al., 2009). The resulting variable-selection problem is thus converted to choosing an appropriate k value.

Screening thresholds admitting an exact low-dimensional oracle property have been studied for various models. With respect to binary classification, screening thresholds for several score functions enabling sure independent screening and exact recovery of active sets have been proposed (Roy et al., 2022). The theoretical properties of those methods, alongside SIS, have moreover been investigated under the framework of the generalized linear models with a binary response (Jiang et al., 2021). Several thresholding methods that offer oracle-error control have also been developed in the context of Gaussian graphical models, in which the presence or absence of a relationship between two nodes corresponds to the value of an entry in the inverse covariance matrix (Jiang et al., 2021; Pan et al., 2017). By employing a doubly-selection estimation procedure, variable selection and valid-error control in high-dimensional discrete-time dynamic systems with a large number of covariates have been achieved while guaranteeing an approximately identical performance both as selection and model fit (Zhao et al., 2020).

Principal Component and Partial Least Squares Methods

Principal component analysis (PCA) and partial least squares (PLS) constitute multivariate linear dimension-reduction techniques widely deployed across domains such as genomics, neuroimaging, biology, environmental science, and engineering. Both approaches facilitate high-dimensional data exploration while safeguarding essential structural information. PCA derives a small number of components linear combinations of original variables that capture predominantly informative content, often requiring retention of only a few components to account for the majority of variance. PLS generalizes this idea to multiple datasets measured on identical units, generating components that maximize covariance between dissimilar data blocks (Liquet et al., 2024). Although PCA and PLS have gained extensive acceptance in the statistical community, the provision of a probabilistic framework remains limited and their interpretative advantages relative to classical regression formulations are not uniformly conferred (Etievant & Viallon, 2020).

Independent Component Analysis and Nonlinear Transformations

Independent component analysis (ICA) seeks to recover independent sources from observed mixtures, preserving the functional relationships between the components. ICA thus strongly addresses the identifiability and interpretability concerns of high-dimensional logistic regression on the data-generating model for which it is designed (Bedychaj et al., 2020). This parallel clarifies extensions of ICA to broader latent-variable models, such as those incorporating nonlinear mixing or time-variant functions All genuine ICA models remain compatible with logit-link observational modeling and Gaussian mixture priors, embedded alongside standard graph-structure modeling (Sasaki et al., 2019). Interaction graph observables, for instance, fully describe every Gaussian-measure ICA model while remaining nonconfounding and readily interpretable, enabling the semiparametric, composite estimation of these high-dimensional, growth-targeted ICA models.

Nonlinear transformations provide a complementary yet related approach to dimensionality reduction. The notion of functional (rather than merely observational) independence underscores, and even defines, the conceptual proximity to ICA. The two objectives diverge in that ICA seeks to recover components as a preprocessing or denoising stage prior to conventional modeling, whereas nonlinear transforms are incorporated directly at the modeling stage. Consequently, ICA estimates the dependent data-generating model at a lower dimensionality, whereas nonlinear approaches target the full-dimensional model.

IV. Modern Machine Learning Frameworks Adapted To P>N

Contemporary statistical and machine-learning classifiers demonstrate a remarkable ability to handle high-dimensional data, yet published theoretical results affirm that classical logistic regression retains its foundational status in this domain. Nevertheless, under even nominally challenging p>n conditions, such non-parametric models still exhibit statistically and computationally efficient p>n adaptation. In practice, however, many modern classifiers tree-based methods, support vector machines, and deep learning thrive in very-high-dimensional settings without explicit p>n adjustments or derivations, leading to uncertainty about their categorization. By contrast, other commonly employed models kernel methods, naïve Bayes, and artificial neural networks are widely considered well-developed for high-dimensional, yet low-information p>n procedures.

High-dimensional penalized-likelihood logistic regression is widely available in software for many popular modelling languages, and it has been implemented in R, Python, and other statistical environments. In general, the elasticity of software implementation for high-dimensional logistic regression matches the high-dimensional, low-information-interdependence conditions of real-world datasets across multiple disciplines. State-of-the-art tools for high-dimensional, high-information Signal-Activation data remain compatible with the p>n actions required by JM apper-simulated p>n test settings. By commonly available modelling platforms rather than proprietary or specialized products are readily accessible for high-dimensional, low-interdependence learning, additional options open for comparable p>n selection.

Prominent frameworks for dealing with high-dimensional data under p>n considerations include penalized-likelihood framework implementations such as coordinate descent algorithms featuring both L1 and L2-regularization support offer high-dimensional logistic regression with performance properties highly competitive with algorithms designed specifically for this application. Well-established implementation through the R package glmnet combines these attractive properties with support across multiple programming environments. Moreover, a wide variety of tree-based classifiers random forests, gradient boosted trees, and extreme gradient boosting handle very-high-dimensional feature sets seamlessly and are supported by proficiency across multiple scripting languages. Regularized tree-based classifiers have been explicitly made available for penalized-likelihood logistic regression (Murris et al., 2022) , although S3-compatible high-dimensional ‘DesignData’ datasets operate consistently within the L0 and Lα standards of a classical p>n-testing same-origin focus. Support vector machines empower the use of explicitly declared high-dimensional features. Following expansion, both kernel-choice and contributor-sparsity specifications become accessible, allowing sparse regimes to accommodate virtually unbounded p parallel to the total count of training samples. Established competence in high-dimensional kernel-support learning permits the maintenance of de facto reference status. Finally, deep-learning adaptations of conventional text-classification architectures carry inherent tenets tailored to fundamental p>n constraints and are well documented in earlier sections.

Penalized Logistic Regression Variants

High-dimensional logistic regression receives growing attention as the statistical community embraces p>n problems in classification. Penalized maximum likelihood provides a standard, widely implemented approach; its variants boast increasing generality, flexibility, and practical reach. The present survey compiles notable refinements across diverse disciplines, systematizes selected enhancements, and outlines related implications. These variants demand increasing computational resources, complicating execution in more stringent operational environments.

Penalized logistic regression emerged as a key approach for handling high-dimensional datasets in practice. Modern formulations extend the foundational theory with new, formalized properties, increasing understanding of their application. All members of this family address established concerns pertaining to stability, interpretability, and scalability. Many, however, address additional challenges arising from domain-specific developments or usage scenarios beyond purely quantitative datasets (Salehi et al., 2019).

Tree-Based Methods with Regularization

Tree-based methods have received significant attention as powerful tools for high-dimensional regression and classification. Despite work on high-dimensional properties of random forests (Deng & Runger, 2012) or boosted trees (Salehi et al., 2019), however, these models have been rarely revisited in substantially

high-dimensional classification settings. Random forests build an ensemble of decision trees trained on bootstrapped data samples. At each node, a random subset of features is considered during branching to encourage diversity. As a result, only a small fraction of variables are involved in any given tree, and many features may remain entirely unselected. Gradient boosting accommodates high-dimensional settings by applying a regularized fit often a decision tree to the residuals of an ensemble model. The regularization term requires additional tuning but introduces sparsity and reduces overfitting. Trees fitted to intermediate logits during the boosting procedure also support high-dimensional coordinate-based feature selection.

Support Vector Machines with High-Dimensional Features

Support vector machines (SVMs) have found broad applications in biological data and Big Data with a sample size smaller than 500. The SVM problem with high-dimensional features retains its 1-norm form, allowing sparsity to be harnessed. The sparsity property can be exploited with embedded structured variable-selection strategies and hierarchical models. The ϵ-insensitive SVM suitable for regression problems can also be extended to selective unformatted input and high-dimension settings. The convex-concave minimization method improves computational efficiency in solving homogeneous polynomial SVM (Wu et al., 2007). SVMs with high-dimensional features are typically implemented by kernel approaches, which map data into high-dimensional feature space while preserving the inner product structure. The choice of kernel function and kernel parameters is critical to the classifier's performance. While many standard kernel functions such as Gaussian, polynomial, and logistic functions exist, the adaptive composite kernel combined with the hyper-parameter optimization mechanism has been developed to improve SVM performance. The adaptive composite kernel is more flexible to construct a good-fit kernel than a fixed kernel of certain type, and it retains the advantage of the simple structure of the existed standard kernels (Weihs & Kassner, 2019).

Deep Learning Adaptations under p>n Constraints

Deep learning adaptations under p>n constraints focus on issues like stability, generalization, and efficient training. Techniques such as stochastic gradient descent stability help train faster and improve generalization. Deep learning methods have been applied in finance, for example, to develop deep portfolios. Large-batch training impacts include potential generalization gaps and sharp minima. Efficient backpropagation techniques optimize neural network training. Recursive error representation methods refine algorithm accuracy. Bayesian perspectives in deep learning facilitate probabilistic modeling. Data augmentation strategies enhance Bayesian deep learning performance. Research also explores deep partial least squares for IV regression, feature selection for personalized policy analysis, and generative causal inference (Polson & Sokolov, 2023).

Applications demand high-dimensional datasets, challenging the assumption that n significantly exceeds p for consistent estimates. Proposals include shallow networks, regularization, early stopping, feature embeddings to map high p to low-dimensional spaces, and pretraining on auxiliary low-dimensional datasets. In optimal transport, high-dimensional distances can be estimated reliably with just a few n points. Pending analysis remains on the instability of gradient-based estimation, potentially counteracted by additive noise or functional regularization (Agarwal et al., 2013).

V. Statistical Properties And Theoretical Insights

High-dimensional logistic regression builds on on classical maximum likelihood estimation (MLE) theory. Under a high-dimensional model representing binary responses determined by latent continuous variables (Sur et al., 2017) , the strategy yields explicit characterizations of the MLE, asymptotic theory for statistical inference, and insights into p>p classification. With a finite number of samples, the convergence of MLE-based point estimates towards their true values is guaranteed, providing a theoretical foundation for the consistency of relevant statistics. High-dimensional datasets brought about fundamental changes in the setting, structure, objectives, and low-dimensional assumptions of statistical theory. Studies have uncovered behaviour under high-dimensional MLE, yet no unified theory explicitly sets precedents or comparison standards regarding relational properties of other estimators. Strategic integration also improves classification performance. Popular logistic regression methods exhibit limitations, including the need for identification and interpretation of joint significance across multiple predictors when examining two or more explanatory variables. Basic properties under high-dimensional logistic regression hold under additional restrictions. Considerable research addresses and challenges the hazard posed by high dimensionality in numerous fields, including weather forecasting, genomics, and finance. Statistical analysis in these domains is commonplace, with several variables influencing either publicly observed elements (for example, stock prices) or individually reported factors (such as medical questions). Nevertheless, the volume of mining and processing presently carried out lies beyond broad comprehension. As a consequence, awareness of, access to, and comprehension of diverse data sources are ever more pressing.

Consistency, sparsity, and oracle properties

The p>n problem refers to the situation when the number of parameters (p) to be estimated exceeds the number of observations (n), which is frequently encountered in modern statistical and machine learning applications. Avoiding the introduction of user-defined methods, it is critical to use established statistical models as the foundation for constructing a classifier. Among the available classifiers, logistic regression is the most widely used and recognized. It can also be regarded as an alternative formulation of two famous models, namely, the standard linear model (the conditional model) and the standard additively separable model (the marginal model) (Lv & Fan, 2009). Extensive empirical comparisons validate the adoption of the logistic regression model as a practical benchmark for high-dimensional binary classification procedure. The highly nonlinear nature of p>n classification naturally leads to a plethora of approaches that regularize either the score function or the kernel map in the batch framework. For supervised dimensionality-reduction classification schemes, significance scores are usually utilized to identify the dimensions which are useful for effective prediction. In the field of computer vision, during pre-processing, the dimensionality of image signals is automatically reduced across a diversity of image categories using non-linear projection that is not sensitive to random noise perturbations. The second role of logistic regression is to benchmark other flexible and nonlinear methods for systematic and objective assessment of relative merits of recent methods (Salehi et al., 2019). The theoretical study of sparsity, oracle, and consistency has attracted significant increasing attention and most common settings (Leeb & M. Poetscher, 2007) are formally addressed.

Information-Theoretic Perspectives

In the p > n setting, the sample complexity for estimating logistic regression models typically grows linearly in the number of variables. Specifically, one can consistently estimate the regression parameters when the logarithm of the model size is sufficiently smaller than the sample size (Sur et al., 2017). In high-dimensional scenarios, jointly estimating the regression coefficients and input features (e.g., ranking the input variables according to their relevance) presents a unique challenge. A positive mutual-information-based metric has been proposed as a feature-ranking criterion under certain assumptions about the high-dimensional-logistic-regression model (Salehi et al., 2019).

Asymptotic Behavior under High-Dimensional Regimes

As already discussed, phase transitions govern the behavior of high-dimensional logistic regression, delineating a regime beyond which the maximum-likelihood estimator (MLE) exhibits non-classical behavior. For the MLE to remain well-defined, such that its continuous limits depend solely on features retaining the characteristics of the original data-generating model, it is necessary for the initial signal-to-noise ratio to exceed a critical threshold (Salehi et al., 2019). Below such critical levels, the amplification of noise prevents consistent estimation of the signal direction, even when the number of observations diverges. In contrast, abundant theoretical investigations have detailed the support recovery, consistency, and asymptotic distribution of high-dimensional linear regression none of which remarked on the relevance of this transition; indeed, the existence and non-explosion cases are adequately addressed even for subcritical settings within that paradigm (Sur & J. Candes, 2018). In contrast, the concurrent behavior of generalized-linear models at finite-dimensional points amid diverging primitives has remained largely unexplored. Turning now to the specific setting of high-dimensional logistic regression, and letting (n) denote the sample size, (p) the number of features, and (k) the dimension of the signal, the profile-likelihood ratio test has been proposed. The related theory, originally formulated in discrete and symmetric contexts, establishes non-asymptotic performance guarantees and advocates for adjustment of conventional machine-learning regularization approaches via squeezing of signal-to-noise ratios, thereby enhancing the extraction of purely structural knowledge. Conceding that, with regards to the settings described, high-dimensional logistic regression models have become increasingly prevalent across countless applicative domains, the presence and impact of a phase transition governing such a purely-linear paradigm raises the proposition that, independent of precisely the duration across the full ((n,p))-space, strictly Gaussian estimators fail to retain the classical M-estimation characteristics anticipated by either the presence of (p gg n) or the escalation of the same within a predominately-regressed Layered Hidden Markov Model.

Practical Considerations and Software Implementations

Considerations for addressing the p>n problem in high-dimensional logistic regression are not limited to methodological choices; they also encompass practical aspects, availability of software, and support for structured workflows. A rich array of tools has been developed to implement high-dimensional logistic regression, facilitating reproducibility and reuse. Hyperparameter tuning constitutes another essential

consideration, acting as a safeguard against overfitting and satisfactory performance. Computational budgets and associated workflow strategies exert a significant influence on method selection. Many scenarios involve progressively addressing multicollinearity, where a substantial collection of candidate predictors undergoes preliminary evaluation before refining the modelling effort. In such workflows, careful consideration of resource allocation plays a crucial role in guiding the choice of either univariate screening, much sought-after by domain experts, or alternative strategies better aligned with p>n challenges. Software implementations for high-dimensional logistic regression methods are diverse yet exhibit certain thematic overlaps, which are relevant both from an implementation standpoint and for offering practical guidance. Various implementations favour a straightforward interface and scalable computation, satisfying fundamental requirements that frequently underpin p>n analysis methods. Select scenarios leverage generative approaches, such as multivariate Gaussian or Dirichlet distributions, to enable efficient high-dimensional proposal generation, in conjunction with extension to category data the influence of compatibility effects remains uncertain.

VI. Systematic Reviews And Meta-Analyses (2015-2025)

Counts of binomial data arise from various fields; many data collections, particularly experiments, contribute information to a binary response. For initial research questions, systematic reviews describe practical procedures for obtaining preliminary qualitative ideas on approaches from the literature already analyzed. Meta-Analyses examine single-proportion and comparison-proportions counts-per-subject datasets; the specified patterns of the already seen total and proportions of the process and the outcomes help consolidate examinations and to gain general ideas on how these and other systematic works will be constructed each fields separately and analyzed and compared each them.

Benchmark Datasets and Evaluation Protocols

The p>n problem is the setting where the number of predicted variables p exceeds the number of observations n, which has become common with the acquisition of high-dimensional data. In this situation, much progress has been made since these methods may present several advantages such as high speed for high-dimensional datasets or the availability of numerous grid-searching tuning-methods for hyper-parameters. Logistic Regression has long been considered a reference classifier for binary classification since it is simple to implement and interpret, and several tools are available in most RFM or ML software. Logistic Regression remains the richest and still one of the most widely used statistical frameworks for dependent-variable analysis. A few comprehensive survey articles or reviews on high-dimension data exist. Most of them focus on protected stability selection, which consists of a systematic, partially automatized selection between the numerous logistic regression approaches. They explore cross-validated filtering to eliminate irrelevant input variables from stable estimation methodologies like selection or regularization. In addition to summarizing the progress, another objective is to identify methodological integration and to provide the practitioner with useful guidance on key problems of methodology selection. A large number of benchmark studies have been published, with the majority comparing two or three methods on five to ten datasets. A recent benchmark study provides insights on the advantages of random forest over logistic regression that have been observed in numerous comparisons. The benchmark study clearly illustrates that the conclusion of a comparative study depends significantly on the properties of the datasets under consideration. For a fair evaluation of protected binary-classification procedures when p>n, a first step is to identify commonly-used datasets and data-split schemes. Typical dataset properties include size, ratio p/n, dimension p, nonlinearity, target-class-cardinality &-imbalance, encoding type, presence of data artifacts, and expected prediction-time-budget. (Couronné et al., 2018).

Reproducibility and Open Science Practices

The p>n problem arises when the number of variables exceeds the number of samples. It is crucial to understand that conventional methods such as classical hypothesis testing, maximum likelihood estimation, and valid confidence intervals are no longer applicable. Logistic regression is widely used to estimate the probability of a binary response from multiple predictor variables, especially in social sciences, finance, and medical sciences. Classical theory states that, with large sample sizes and fixed predictors, the maximum likelihood estimator follows an asymptotic normal distribution, justifying the use of standard errors and statistical tests. Estimates obtained under low-dimensionality assumptions continue to play a role in the analysis of high-dimensional data. The hypothesis that the features still determine the response remains relevant in many situations. Despite the pervasiveness of high-dimensional data, research articles on quantitative social science topics rarely include analyses of high-dimensional datasets. The absence of high-dimensional empirical analyses has prevented the identification of topics that could be targeted for further methodological development and consideration in quantitative studies.

Parameter estimation involves the selection of a set of relevant variables that contribute to the prediction of the target variable and the estimation of the corresponding coefficients. Properly defined target

variables include continuous variables and categorical variables with two or more classes. Methods used for predicting continuous variables are text mining, internet browsing data, and other multivariate time series techniques. The ability to uncover hidden structure in these types of high-dimensional data is required. High-dimensional classification deals with selecting features and building classification models. Reproducibility and open science practices are essential to addressing the replicability crisis (Vsevolozhskaya et al., 2017). Prediction intervals provide researchers with a quantitative assessment of what to expect in repeated analyses using independent samples. This approach helps evaluate variability inherent in statistics and P-values, supporting the robustness of findings. The state of play of reproducibility analysis across ten statistical disciplines and the tracking of reproducibility initiatives and indicators are conducted (Xiong & Cribben, 2022). Reproducibility efforts include promoting independent analysis of industry-sponsored studies and creating multi-language computing environments for literate programming. The phyloseq project provides an open-source R tool for reproducible analysis of phylogenetic sequencing data. Addressing reproducibility issues, initiatives have highlighted the lack of supporting code and data in many statistical publications. Researchers have proposed frameworks like research compendiums for organizing digital research materials, enabling others to reproduce and extend research. R packages have been developed for tracking, annotating, and reproducing computational results. Addressing statistical challenges like selective inference is seen as key for enhancing replicability. Tools and standards are being created to prevent and analyze reproducibility errors in data science projects, aiming to improve the reliability of published research. Logistic regression is a fundamental model learned early in statistics education, with standard interpretations of output like regression coefficients, standard errors, and p-values. In modern data analysis, huge datasets with many features challenge traditional approximations, raising questions about their validity when the number of predictors is large relative to the sample size (Sur & J. Candes, 2018).

VII. Case Studies Across Disciplines

High-dimensional datasets naturally occur in many real-life applications, including, spectrometry, satellite imagery, and microarrays. Despite efforts to gather information through feasible experimental designs and rigorous protocols, human interests subject such data to various constraints. These validations confirm patterns addressed at the data-gathering stage, focusing on aspects of substance or knowledge that are not sterile and trivial routine checkpoints. Analyses related to public health, medicine, economics, and socio-demographic examine significant motivational determinants such as gender, religion, and age by analyzing outcome distribution. Social and behavioral sciences further broaden the scope of domain problems. Attention to prior consultation detects serious experimental flaws that hinder significance. Strong algorithms in such domains often face unanticipated obstacles, as irrelevant variables and artifacts inadvertently enter the market.

Genomics and Proteomics

Genomic, proteomic, and related biochemical datasets present extremely high-dimensional responses, yet variable selection and subsequent classification remain pressing issues. Initial screening of sequentially collected gene-expression data enables the construction of relevant biological models for patients suffering from well-characterized diseases, such as specific cohorts of cancer patients (Afsari et al., 2014). The biological plausibility of models remains essential, as additional experiments and therapeutic strategies depend on biological validation of the selected features and the model itself. Prognosis or prediction following large-scale experiments, such as those involving mRNA or protein samples, often invoke a second generation of data acquired on a different set of patients. In these cases, biomarker discovery aims to uncover a small number of informative genes or proteins that can be validated and potentially employed for therapeutic monitoring across a clinically wide cohort of cases. Examples include serum cytokines or RNA from blood cells, where “wider” models are typically preferred, since the objective still concerns the discovery of a biological signature on the starting “big dataset” (Bazzoli & Lambert-Lacroix, 2018).

Biomedical Imaging

High-dimensional biomedical imaging datasets have seen explosive growth, yet most analysis methods fail to scale sufficiently. Even when only segmentations are considered, high-dimensional methods are still necessary. Certain high-dimensional techniques also provide a predictive perspective that is distinct from feature selection. Computational times appear to scale favorably alongside the dimension. Direct interpretability for image covariates generally remains limited; more accessible interpretation is feasible when models focus instead on other features. The extensive literature on high-dimensional techniques in these fields presents many opportunities for and obstacles to multinomial cross-fertilization (An & Zhang, 2020).

Social and Behavioral Sciences

Segmentation analysis is widely applied in behavioral and social sciences. Survey-based segmentation aims to identify respondents who desire similar interventions based on shared views or actions. Classification of psychometric types, such as the Big Five, helps in personality-based marketing strategies. Research in educational psychology often concerns the classification of cognitive and metacognitive learning strategies. Others attempt to classify online customer profiles based on internet usage behavior. Applying behavioral data to human behavior prediction has become increasingly popular. Logistic regression can be adapted to classify individuals on selected behavioral characteristics, a task often pursued with fine-tuning and Gaussian mixture model seed methods based on pre-trained BERT embeddings. Regularized multinomial logistic regression is employed to detect psychological symptom patterns from high-dimensional Twitter data, with a Bayesian prior that combines sparsity and commonality, achieving significant model selection and covariate-recovery capabilities in state-of-the-art performance. Sparse principal covariate regression, a unifying bridge between principal component analysis and sparse estimation, identifies the most essential heterogeneous patterns from all 112,000 Twitter feeds released by the 2020 COVID-19 campaign in France. Data fusion models can carefully extract useful knowledge from various high-dimensional datasets, such as users' Twitter feeds, online news, online videos, and PDF files, to provide comprehensive reporting on the support of mental wellbeing or mental illness.

Gaps, Controversies, and Future Directions

The most impactful challenges in high-dimensional data analysis derive from a fundamental tension between aspirations for more informative models and the recognition that classical approaches fail and require substantial modification to become viable in the high-dimensional context (Sur & J. Candes, 2018). Many of the most widely used models themselves have been deeply modified to cope with new, more modest goals. In this spirit, statistical methods that support the interpretation of classification decisions have been developed for supportive data generation processes where the classification of a target variable depends of counts from auxiliary rogate covariates.

VIII. Conclusion

The statistical preeminence of high-dimensional data and the accompanying p>n problem have rendered logistic regression increasingly popular for binary-response modeling. An evidence-guided evaluation of p>n approaches must therefore begin with a clear understanding of high-dimensional logistic regression, its foundational principles, and the specific challenges presented by a p>n framework. While untangling methodological critique lodged against p>n logistic regression, it also becomes apparent that high-dimensional applications impose strict requirements on practitioners. In particular, the suitability of all tested p>n methods hinge on clearly defined datasets and associated performance metrics, suggesting that accompanying numerical illustrations should elucidate data-generation processes and interpretation.

High-Dimensional Binary Regression (Sur & J. Candes, 2018). Logistic regression is the predominant model for estimating the probability of a binary response as a function of multiple predictors, with broad application across the social, financial, and medical sciences. It ranks among the first topics covered in statistical and data-analysis courses due to its critical role in interpreting key outputs such as regression coefficients, standard errors, and p-values. Classical theory establishes that for large sample sizes, with the number of predictors held fixed, the maximum-likelihood estimator (MLE) of the coefficients is approximately normally distributed, thus enabling valid inference and significance testing. Wilks' theorem further identifies the asymptotic distribution of the likelihood-ratio test, which similarly undergirds statistical testing. Yet the continued validity of these classical approximations remains unclear in modern analysis involving extremely large datasets with high-dimensional features, in which the number of predictors becomes comparable to the sample size.

References

[1]. Sur, P., & Candès, E. J. (2019). A Modern Maximum-Likelihood Theory For High-Dimensional Logistic Regression. Proceedings Of The National Academy Of Sciences, 116(29), 14516-14525.
[2]. Sur, P., Chen, Y., & Candès, E. J. (2019). The Likelihood Ratio Test In High-Dimensional Logistic Regression Is Asymptotically A Rescaled Chi-Square. Probability Theory And Related Fields, 175(1), 487-558.
[3]. Salehi, F., Abbasi, E., & Hassibi, B. (2019). The Impact Of Regularization On High-Dimensional Logistic Regression. Advances In Neural Information Processing Systems, 32.
[4]. An, B., & Zhang, B. (2020). Logistic Regression With Image Covariates Via The Combination Of L 1 And Sobolev Regularizations. Plos One, 15(6), E0234975.
[5]. Liu, Z. (2017). An Aggregation Method For Sparse Logistic Regression. International Journal Of Data Mining And Bioinformatics, 17(1), 85-96.
[6]. Kurnaz, F. S., Hoffmann, I., & Filzmoser, P. (2018). Robust And Sparse Estimation Methods For High-Dimensional Linear And Logistic Regression. Chemometrics And Intelligent Laboratory Systems, 172, 211-222.

[7]. Balcan, M. F. F., Khodak, M., Sharma, D., & Talwalkar, A. (2022). Provably Tuning The Elasticnet Across Instances. Advances In Neural Information Processing Systems, 35, 27769-27782.
[8]. Chakraborty, A. (2018). Bayesian Shrinkage: Computation, Methods And Theory (Doctoral Dissertation, Texas A&M University).
[9]. Fang, Y., Wang, J., & Sun, W. (2013). A Note On Selection Stability: Combining Stability And Prediction. Arxiv Preprint Arxiv:1301.7118.
[10]. Pokarowski, P., Mielniczuk, J., & Teisseyre, P. (2012). Linear Regression Model Selection Using P-Values When The Model Dimension Grows. Arxiv Preprint Arxiv:1205.4146.
[11]. Liu, X. (2015). On Testing Common Indices For Several Multi-Index Models: A Link-Free Approach.
[12]. Weng, J., & Young, D. S. (2017). Some Dimension Reduction Strategies For The Analysis Of Survey Data. Journal Of Big Data, 4(1), 43.
[13]. Roy, S., Sarkar, S., Dutta, S., & Ghosh, A. K. (2024). On Exact Feature Screening In Ultrahigh-Dimensional Binary Classification. Journal Of Computational And Graphical Statistics, 33(2), 448-462.
[14]. Liquet, B., Moka, S., & Muller, S. (2025). Best Subset Solution Path For Linear Dimension Reduction Models Using Continuous Optimization. Biometrical Journal, 67(1), E70015.
[15]. Etiévant, L., & Viallon, V. (2022). On Some Limitations Of Probabilistic Models For Dimension Reduction: Illustration In The Case Of Probabilistic Formulations Of Partial Least Squares. Statistica Neerlandica, 76(3), 331-346.
[16]. Bedychaj, A., Spurek, P., Nowak, A., & Tabor, J. (2020). WICA: Nonlinear Weighted ICA. Arxiv Preprint Arxiv:2001.04147.
[17]. Sasaki, H., Takenouchi, T., Monti, R., & Hyvarinen, A. (2020, August). Robust Contrastive Learning And Nonlinear ICA In The Presence Of Outliers. In Conference On Uncertainty In Artificial Intelligence (Pp. 659-668). PMLR.
[18]. Murris, J., Charles-Nelson, A., Lavenu, A., & Katsahian, S. (2022). Towards Filling The Gaps Around Recurrent Events In High-Dimensional Framework: Literature Review And Early Comparison. Arxiv Preprint Arxiv:2203.15694.
[19]. Deng, H., & Runger, G. (2012, June). Feature Selection Via Regularized Trees. In The 2012 International Joint Conference On Neural Networks (IJCNN) (Pp. 1-8). IEEE.
[20]. Wu, S., Zou, H., & Yuan, M. (2008). Structured Variable Selection In Support Vector Machines.
[21]. Polson, N., & Sokolov, V. (2023). Deep Learning: A Tutorial. Arxiv Preprint Arxiv:2310.06251.
[22]. Weihs, C., & Kassner, T. (2018). Classification Method Performance In High Dimensions. Universitätsbibliothek Dortmund.
[23]. Agarwal, A., Kakade, S., Karampatziakis, N., Song, L., & Valiant, G. (2014, June). Least Squares Revisited: Scalable Approaches For Multi-Class Prediction. In International Conference On Machine Learning (Pp. 541-549). PMLR.
[24]. Lv, J., & Fan, Y. (2009). A Unified Approach To Model Selection And Sparse Recovery Using Regularized Least Squares.
[25]. Leeb, H., & Pötscher, B. M. (2008). Sparse Estimators And The Oracle Property, Or The Return Of Hodges' Estimator. Journal Of Econometrics, 142(1), 201-211.
[26]. Couronné, R., Probst, P., & Boulesteix, A. L. (2018). Random Forest Versus Logistic Regression: A Large-Scale Benchmark Experiment. BMC Bioinformatics, 19(1), 270.
[27]. Vsevolozhskaya, O., Ruiz, G., & Zaykin, D. (2017). Bayesian Prediction Intervals For Assessing P-Value Variability In Prospective Replication Studies. Translational Psychiatry, 7(12), 1271.
[28]. Xiong, X., & Cribben, I. (2023). The State Of Play Of Reproducibility In Statistics: An Empirical Analysis. The American Statistician, 77(2), 115-126.
[29]. Afsari, B., Braga-Neto, U. M., & Geman, D. (2014). Rank Discriminants For Predicting Phenotypes From RNA Expression.
[30]. Bazzoli, C., & Lambert-Lacroix, S. (2018). Classification Based On Extensions Of LS-PLS Using Logistic Regression: Application To Clinical And Multiple Genomic Data. BMC Bioinformatics, 19(1), 314.