STATISTICAL LEARNING WITH EMPIRICAL FEATURES AND DATA OF DIFFERENT TYPES

HUIHUI QIN

PhD

The Hong Kong Polytechnic University

2020

The Hong Kong Polytechnic University

Department of Applied Mathematics

Statistical Learning with Empirical Features and Data of Different Types

Huihui Qin

A thesis submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy

August 2020

Certificate of Originality

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

____________________________(Signed)

_________QIN Huihui_________(Name of student)

Abstract

The thesis consists of three parts that cover different aspects of statistical learning for data mining.

In the first part, we propose a new algorithm, LESS (Learning with Empirical feature-based Summary statistics from Semi-supervised data), which uses only summary statistics instead of raw data for regression learning. Nowadays the extensive collection and analyzing of data is stimulating widespread privacy concerns, and therefore is increasing tensions between the potential sources of data and researchers. A privacy-friendly learning framework can help to ease the tensions, and to free up more data for research. In LESS, The selection of empirical features serves as a trade-off between prediction precision and the protection of privacy. We show that LESS achieves the minimax optimal rate of convergence, in terms of the size of the labeled sample. LESS extends naturally to the applications where data are separately held by different sources. Compared with existing literature on distributed learning, LESS removes the restriction of minimum sample size on single data sources.

In the second part of the thesis, we study different approaches for analyzing topics in text data. Topic modeling has been an important field in natural language processing (NLP) and recently witnessed great methodological advances. Yet, the development of topic modeling is still, if not increasingly, challenged by two critical issues. First, despite intense efforts toward nonparametric/post-training methods, the search for the optimal number of topics K remains a fundamental question in topic modeling and warrants input from domain experts. Second, with the development of more sophisticated models, topic modeling is now ironically been treated as a black box and it becomes increasingly difficult to tell how research findings are informed by data, model specifications, or inference algorithms. Based on about 120,000 newspaper articles retrieved from three major Canadian newspapers (Globe and Mail, Toronto Star, and National Post) since 1977, we employ five methods with different model specifications and inference algorithms (Latent Semantic Analysis, Latent Dirichlet Allocation, Principal Component Analysis, Factor Analysis, Non- negative Matrix Factorization) to identify discussion topics. The optimal topics are then assessed using three measures: coherence statistics, held-out likelihood (loss), and graph-based dimensionality selection. Mixed findings from this research complement advances in topic modeling and provide insights into the choice of optimal topics in social science research.

In the third part, we consider the generalized linear hurdle model with grouped and right-censored count data. This data type is widely applied in demography, epidemiology, sociology, criminology, psychology, and many other branches of social sciences. The corresponding generalized linear model and the zero-inflated model recently draw much attention. In this part, we study the hurdle model which covers not only zero inflation but also zero deflation. We provide sufficient conditions for the asymptotic consistency and asymptotic normality of maximum likelihood estimator. We represent the Fisher information matrix of the hurdle model in terms of the vanilla grouped and right-censored model. We provide an elegant sufficient and necessary condition for the Fisher information matrix of the hurdle model to be strictly positive definite. The research complements the recent development of the statistical inference with grouped and right-censored count data.

Acknowledgements

I express my sincere acknowledgment to my two supervisors, Professor Jian Huang and Dr. Xin Guo. Professor Jian Huang led me to the academic world. His novel points and helpful discussions benefit me a lot. Dr. Xin Guo led me to the learning theory world. His patient guidance has a profound influence on me.

I thank the Hong Kong Polytechnic University and the Department of Applied Mathematics. In this beautiful institute, I enjoyed precious resources, which assisted my research in the past three years. I thank all the staff and classmates who gave me help in the past.

I give my thanks to my family. I thank my parents for giving me life and always supporting me in the back. They always encouraged me to chase my dream in the past 20 years. I will forever treasure their love and support!

Contents

Chapter 1

Semi-supervised Learning with Summary Statistics

1.1 Introduction

Many reproducing kernel-based machine learning algorithms are designed without considering privacy issues. In particular, under the structural risk minimization scheme, as pointed out by the representer theorem, the whole input part of training data, which may contain private information, has to be shipped along with the predicted function. Privacy concern would restrict the application of such algorithms. On the other hand, usually there are unlabeled data available with the same marginal distribution as the training data. For example, these unlabeled data could be produced by sampling from the estimated density, or be obtained from public domain without privacy issues [105, 66]. In this paper, we study the methodology for masking the sensitive private information in training data, with the help of unlabeled data.

Semi-supervised learning is a big class of machine learning problems where unlabeled data are used in addition to the data points with labels, e.g., for classification or regression. In recent years, unlabeled data are observed helpful for capturing the underlying manifold structures of data distribution [21, 8], relaxing the requirement on single-source minimum sample size in distributed learning [64, 45], and improving the convergence under weak regularity assumptions of the regression function [45].

In this chapter, unlabeled data (possibly also including the input part of the labeled data) are used to build empirical features first. Then, we use the empirical features to construct summary statistics, based on which we introduce a new algorithm, **LESS** (Learning with Empirical feature-based Summary statistics from Semi-supervised data), of which the main advantages are summarized below.

- LESS achieves the minimax optimal convergence rate, in terms of the size of labeled sample.

- With the help of unlabeled data, LESS has an automatic generalization to distributed learning, where the restriction on single-source minimum sample size is completely removed.

- The summary statistics we adopt provide a protocol for communicating data with privacy. Unlike classical kernel-based algorithms, LESS collects only the summary statistics, instead of the private raw data, for the centralized learning process.

Consider a regression learning problem with an input space X, which is a compact metric space, and an output space $Y \subset \mathbb{R}$. Let $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ be a sample drawn independently from $(Z = X \times Y, \rho)$, where ρ is an unknown Borel probability measure such that the marginal distribution ρ_X on X is nondegenerate, i.e., $\rho_X(A) > 0$ for any measurable set A that has an interior point. The target of the regression problem is to learn the regression function $f_\rho : X \to \mathbb{R}$,

$$f_\rho(x) = \int_Y y d\rho(y|x),$$

from the sample $\mathbf{z}$, where $\rho(y|x)$ is the conditional distribution of ρ at x.

There is a large literature of the kernel methods for machine learning. See [91, 90, 96, 103, 89, 63], and the reference therein. Let $K : X \times X \to \mathbb{R}$ be a Mercer kernel.

That is, K is a function which is symmetric, continuous, and positive, where positive means $\sum_{i,j=1}^{l} c_i c_j K(u_i, u_j) \geqslant 0$ for any integer $1 \leqslant l < \infty$, any coefficients $c_1, \ldots, c_l \in \mathbb{R}$, and any elements $u_1, \ldots, u_l \in X$. Let $(\mathcal{H}_K, \langle \cdot, \cdot \rangle_K, \|\cdot\|_K)$ be the reproducing kernel Hilbert space generated by K. The classical kernel-based regularized least squares algorithm is defined by

$$f_\lambda^{\mathbf{z}} = \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2 + \lambda \|f\|_K^2 \right\}, \tag{1.1}$$

where $\lambda > 0$ is the regularization parameter. Kernel-based learning algorithms usually have the flaws in privacy protection. For example, by the well-known representer theorem [96], $f_\lambda^{\mathbf{z}}$ in (1.1) takes the form

$$f_\lambda^{\mathbf{z}} = \sum_{i=1}^{m} c_i K_{x_i}, \tag{1.2}$$

where $c_1, \ldots, c_m \in \mathbb{R}$ are the coefficients determined by (1.1), and for any $x, u \in X$, the function $K_x : X \to \mathbb{R}$ is defined by $K_x(u) = K(x, u)$. It is easy to see that to ship $f_\lambda^{\mathbf{z}}$, the unlabeled part $\mathbf{x} = \{x_i\}_{i=1}^m$ of the sample $\mathbf{z}$ must be shipped together. We put a discussion in Section 1.3. In this paper, we try to solve this problem on privacy, by introducing the empirical feature-based summary statistics.

We assume that there is another sample $\mathbf{u} = \{u_i\}_{i=1}^n$, drawn independently from ρ_X without labels. For applications, the sample $\mathbf{u}$ may come from some openly accessible sources, for example those with the privacy expired. Note that we do not assume independence between $\mathbf{u}$ and $\mathbf{x}$. In particular, a part, or even the whole of $\mathbf{x}$ could just be put into $\mathbf{u}$. This inclusion is sometimes useful, and is covered by our analysis.

Define $L_K^{\mathbf{u}} : \mathcal{H}_K \to \mathcal{H}_K$ as an operator by

$$L_K^{\mathbf{u}} f = \frac{1}{n} \sum_{i=1}^{n} f(u_i) K_{u_i}, \tag{1.3}$$

where $|\mathbf{u}| = n$ is the size of $\mathbf{u}$. By the reproducing property [22] that for any $f \in \mathcal{H}_K$ and $u \in X$, $\langle f, K_u \rangle_K = f(u)$, one has that for any $f, g \in \mathcal{H}_K$,

$$\langle L_K^{\mathbf{u}} f, g \rangle_K = \frac{1}{n} \sum_{i=1}^{n} f(u_i) g(u_i) = \langle f, L_K^{\mathbf{u}} g \rangle_K .$$

In particular, $\langle L_K^{\mathbf{u}} f, f \rangle_K = \frac{1}{n} \sum_{i=1}^{n} f(u_i)^2 \geqslant 0$. So $L_K^{\mathbf{u}}$ is a positive semi-definite operator with rank (i.e., the dimension of its image) at most n. Therefore, we can write $\{(\lambda_i^{\mathbf{u}}, \phi_i^{\mathbf{u}})\}_i$ as the eigensystem of $L_K^{\mathbf{u}}$ with $\lambda_1^{\mathbf{u}} \geqslant \lambda_2^{\mathbf{u}} \geqslant \ldots \geqslant \lambda_n^{\mathbf{u}} \geqslant 0 = \lambda_{n+1}^{\mathbf{u}} = \ldots$. The zero eigenvalues are counted purposely to make $\{\phi_i^{\mathbf{u}}\}_i$ an orthonormal basis of $\mathcal{H}_K$. Similarly, we define $L_K^{\mathbf{x}}$ and $\{(\lambda_i^{\mathbf{x}}, \phi_i^{\mathbf{x}})\}_i$ for the input part $\mathbf{x}$ of the sample $\mathbf{z}$ by substituting $\mathbf{u}$ with $\mathbf{x}$, and n with $m = |\mathbf{x}|$ in (1.3).

Algorithm LESS. The sample dependent functions $\phi_i^{\mathbf{u}}$'s are referred to as empirical features (so are $\phi_i^{\mathbf{x}}$'s). These functions are studied in literature [46, 110, 111] as powerful tools for regression, classification, and nonlinear dimension reduction. Let $1 \leqslant N \leqslant n$ be an integer. Consider the summary statistic $\mathbf{d} = (d_1, \ldots, d_N)^T$, defined by

$$d_i = \left\langle \phi_i^{\mathbf{u}}, \frac{1}{m} \sum_{j=1}^{m} y_j K_{x_j} \right\rangle_K , \quad 1 \leqslant i \leqslant N. \tag{1.4}$$

The superscripts $\mathbf{u}$ and $\mathbf{z}$ of $\mathbf{d}$ and d_i's are dropped to avoid heavy notation. The summary statistic $\mathbf{d}$ is then used to build the output function of LESS,

$$f_\lambda^{\mathbf{u},\mathbf{z}} = (L_K^{\mathbf{u}} + \lambda I)^{-1} \sum_{i=1}^{N} d_i \phi_i^{\mathbf{u}} = \sum_{i=1}^{N} \frac{d_i}{\lambda_i^{\mathbf{u}} + \lambda} \phi_i^{\mathbf{u}}, \tag{1.5}$$

where $\lambda > 0$ is the regularization parameter, and in this paper, I denotes the identity operator, with its domain inferred from the context. Here, recall that $\phi_i^{\mathbf{u}}$ is an eigenfunction of $L_K^{\mathbf{u}}$, $L_K^{\mathbf{u}} \phi_i^{\mathbf{u}} = \lambda_i^{\mathbf{u}} \phi_i^{\mathbf{u}}$. We have $(L_K^{\mathbf{u}} + \lambda I)^{-1} \phi_i^{\mathbf{u}} = \frac{1}{\lambda_i^{\mathbf{u}} + \lambda} \phi_i^{\mathbf{u}}$.

We see that by the introduction of the empirical features $\phi_i^{\mathbf{u}}$'s, the training sample $\mathbf{z}$ is encoded into $\mathbf{d}$, instead of directly shipped along the predicted function $f_\lambda^{\mathbf{u},\mathbf{z}}$.

From the statistic $\mathbf{d}$, it is even not trivial to recover the sample size m! Of course, a safer design could be achieved by adding noise to $\mathbf{d}$, which we leave as future work.

LESS for distributed learning. The summary statistics $\mathbf{d}$ provides an automatic and unified way for distributed learning. In fact, suppose that instead of (1.4), the sample $\mathbf{z}$ is stored separately in ℓ sources $\mathbf{z} = \mathbf{z}_1 \cup \mathbf{z}_2 \cup \ldots \cup \mathbf{z}_\ell$ without overlapping, then one defines $\mathbf{d}^J = (d_1^J, \ldots, d_N^J)^T$ by

$$d_i^J = \left\langle \phi_i^{\mathbf{u}}, \frac{1}{|\mathbf{z}_J|} \sum_{(x,y)\in\mathbf{z}_J} yK_x \right\rangle_K, \quad 1 \leqslant J \leqslant \ell, \quad 1 \leqslant i \leqslant N. \tag{1.6}$$

Again, one may centralize the summary statistics $\mathbf{d}^J$'s without directly collecting the private data sets $\mathbf{z}_J$'s. More importantly, the weighted average of $\mathbf{d}^J$'s is exactly $\mathbf{d}$,

$$\mathbf{d} = \sum_{J=1}^{\ell} \frac{|\mathbf{z}_J|}{|\mathbf{z}|} \mathbf{d}^J. \tag{1.7}$$

So, without any configuration, LESS can be directly applied to distributed learning problems, where data are separately held by different sources as privacy. From (1.7), we see that the sizes of different data subsets have no effect on the learning process (1.5). In another way of saying, our analysis on LESS applies automatically to this distributed design (1.6).

The rest of this chapter is organized as follows. We first give our main results in Section 1.2. Comparisons and discussions, as well as the details of implementations are put in Section 1.3. Proofs are placed in Section 1.4.

1.2 Main Results

In this section, we formulate the main assumptions and our main results.

Write $(L^2_{\rho_X}, \|\cdot\|_\rho)$ the Hilbert space of square-integrable functions on X with respect to the measure ρ_X. Define $L_K : L^2_{\rho_X} \to L^2_{\rho_X}$ by

$$f \mapsto \int_X f(x) K_x d\rho_X(x).$$

Since K is continuous and X is compact, L_K is compact. It is easy to verify that L_K is positive semi-definite. Furthermore, L_K is of trace class (hence Hilbert-Schmidt), and since ρ_X is nondegenerate, $\|L_K^{1/2} f\|_K = \|f\|_\rho$ for any $f \in L^2_{\rho_X}$. Denote $\kappa = \max\{1, \sup_{x \in X} \sqrt{K(x,x)}\}$. We have $\mathsf{Trace}(L_K) \leqslant \kappa^2$. See [22] for detailed proofs. So we write

$$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant 0,$$

as all the eigenvalues of L_K, and $\phi_1, \phi_2, \ldots$ the corresponding eigenfunctions, normalized in $\mathcal{H}_K$. For $\lambda > 0$, write

$$\mathcal{N}(\lambda) = \mathsf{Trace}(L_K(L_K + \lambda I)^{-1})$$

the effective dimension of L_K [102, 17, 12]. The following assumption (A1) characterizes the capacity of the hypothesis space $\mathcal{H}_K$, and is widely adopted in learning theory literature [64, 63, 11].

(A1) *There exist some constants $0 < C_1 < \infty$ and $0 < s \leqslant 1$ such that $\mathcal{N}(\lambda) \leqslant C_1 \lambda^{-s}$ for any $0 < \lambda < \infty$.*

The following assumption (A2) characterizes the regularity of the regression function.

(A2) *There exists some $g_\rho \in L^2_{\rho_X}$ and $1/2 \leqslant r \leqslant 1$ such that $f_\rho = L_K^r g_\rho$.*

Note that Assumption (A2) implies $f_\rho \in \mathcal{H}_K$.

(A3) *$\int_Z y^2 d\rho(x,y) < \infty$, and that there exist two constants $0 < \sigma, M < \infty$, such that*

$$\int_Y \left(\exp\left\{ \frac{|y - f_\rho(x)|}{M} \right\} - \frac{|y - f_\rho(x)|}{M} - 1 \right) d\rho(y|x) \leqslant \frac{\sigma^2}{2M^2},$$

for ρ_X-almost all $x \in X$.

In particular, (A3) holds with $\sigma = \sqrt{2(e^2 - 3)}M$, when $|y| \leqslant M$ almost surely. For more discussions on (A3), see [64, 7, 17, 97].

From the design (1.4) and (1.5), we see that intuitively, one needs sufficient coordinates for $\mathbf{d}$ to guarantee the convergence. In particular, we characterize the requirement by the following assumption (A4).

(A4) *N is large enough (meaning that enough empirical features are used), so that $\lambda_{N+1} \leqslant \kappa^2 \lambda$.*

Theorem 1.1. *Assume (A1), (A2), (A3), and $n \geqslant m$. For any $0 < \delta < 1$, one has with confidence at least $1 - \delta$ that*

$$\begin{aligned} \|f_\lambda^{\mathbf{u},\mathbf{z}} - f_\rho\|_\rho \leqslant & \left(\frac{2\mathcal{B}_{n,\lambda}^2}{\lambda} + 2 \right) \left(\frac{M + \sigma}{\kappa} + \|f_\rho\|_K + \|g_\rho\|_\rho \right) \mathcal{B}_{m,\lambda} \log^3 \frac{10}{\delta} \\ & + \left(\frac{2\mathcal{B}_{n,\lambda}^2}{\lambda} + 2 \right)^r \left(\lambda + \frac{4\kappa^2}{\sqrt{n}} + \lambda_{N+1} \right)^r \|g_\rho\|_\rho \log^{3r} \frac{10}{\delta} + \|g_\rho\|_\rho \lambda^r, \end{aligned}$$

where

$$\mathcal{B}_{n,\lambda} = \frac{2\kappa^2}{n\sqrt{\lambda}} + 2\kappa \sqrt{\frac{\mathcal{N}(\lambda)}{n}}, \tag{1.8}$$

and $\mathcal{B}_{m,\lambda}$ is similarly defined by substituting n with m.

We cite from [45, Lemma B.1] the following lemma, which is standard, and the proof can also be found in [63] and [48, Lemma 11].

Lemma 1.1. *Let R be a nonnegative random variable. Let $\alpha, \beta, \gamma > 0$. If for any $0 < \delta < 1$, one has with confidence at least $1 - \delta$ that $R \leqslant \alpha \log^{\gamma} \frac{\beta}{\delta}$, then for any $\mu > 0$,*

$$(\mathbb{E}[R^{\mu}])^{1/\mu} \leqslant \alpha \left[\beta \Gamma(\mu\gamma + 1)\right]^{1/\mu},$$

where $\Gamma(t) = \int_0^{\infty} e^{-u} u^{t-1} du$ is the Gamma function.

Corollary 1.1. *Assume (A1), (A2), (A3), (A4), and $n \geqslant \max\{m, m^{\frac{2}{2r+s}}\}$. Let $\lambda = m^{-\frac{1}{2r+s}}$. For any $0 < \delta < 1$, one has with confidence at least $1 - \delta$ that*

$$\|f_{\lambda}^{\mathbf{u},\mathbf{z}} - f_{\rho}\|_{\rho} \leqslant C_2 m^{-\frac{r}{2r+s}} \log^3 \frac{10}{\delta}, \tag{1.9}$$

where C_2 is a constant independent of m, n, or δ, and it is given at the end of the proof. Moreover, for any $\mu > 0$, Lemma 1.1 gives

$$\left[\mathbb{E}(\|f_{\lambda}^{\mathbf{u},\mathbf{z}} - f_{\rho}\|_{\rho}^{\mu})\right]^{1/\mu} \leqslant C_2 \left[10\Gamma(3\mu + 1)\right]^{1/\mu} m^{-\frac{r}{2r+s}}. \tag{1.10}$$

Remark 1.1. *Recall that $1 \leqslant N \leqslant n$. With the assumption $n \geqslant \max\{m, m^{\frac{2}{2r+s}}\}$ and the setting $\lambda = m^{-\frac{1}{2r+s}}$, it is always possible to find some $N \leqslant n$ that satisfies Assumption (A4). In fact, since the eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \ldots$ of L_K are arranged in non-increasing order, $\lambda_n \leqslant \frac{1}{n}\mathsf{Trace}(L_K) \leqslant \frac{\kappa^2}{n} \leqslant \kappa^2 m^{-\frac{1}{2r+s}} = \kappa^2 \lambda$.*

Remark 1.2. *It is well understood [17, 92, 7] that when $1/2 \leqslant r \leqslant 1$, the minimax optimal learning rate for learning algorithms that have only the access to $\mathbf{z}$ and with output functions in $\mathcal{H}_K$, is $O(m^{-\frac{r}{2r+s}})$. The bounds (1.9) and (1.10) in Corollary 1.1 match this rate.*

1.3 Discussions and Comparisons

1.3.1 Details for the Implementations

Recall $m = |\mathbf{x}|$. Define the sampling operator $S_{\mathbf{x}} : \mathcal{H}_K \to \mathbb{R}^m$,

$$f \mapsto (f(x_i))_{i=1}^m .$$

It is straightforward to see that the adjoint operator $S_{\mathbf{x}}^T : \mathbb{R}^m \to \mathcal{H}_K$ is defined by

$$(c_i)_{i=1}^m \mapsto \sum_{i=1}^m c_i K_{x_i}.$$

Let $\mathbb{K}$ be the Gram matrix of the Mercer kernel K on $\mathbf{x}$, $\mathbb{K} = (K(x_i, x_j))_{i,j=1}^m$. Then,

$$\frac{1}{m}\mathbb{K} = \frac{1}{m} S_{\mathbf{x}} S_{\mathbf{x}}^T, \qquad L_K^{\mathbf{x}} = \frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}}. \tag{1.11}$$

So the eigenvalues of $\frac{1}{m}\mathbb{K}$, counting multiplicity, are $\lambda_1^{\mathbf{x}}, \ldots, \lambda_m^{\mathbf{x}}$, which are the first m eigenvalues of $L_K^{\mathbf{x}}$. Since $\mathbb{K}$ is positive semi-definite, we have the following eigendecomposition

$$\frac{1}{m}\mathbb{K} = U \Lambda U^T, \qquad \Lambda = \mathrm{diag}\{\lambda_1^{\mathbf{x}}, \ldots, \lambda_m^{\mathbf{x}}\},$$

where $U = [U_1, \ldots, U_m]$ is an orthogonal matrix. Some simple linear algebra shows that if $\lambda_i^{\mathbf{x}} = 0$, then $\langle \phi_i^{\mathbf{x}}, L_K^{\mathbf{x}} \phi_i^{\mathbf{x}} \rangle_K = 0$, so $S_{\mathbf{x}} \phi_i^{\mathbf{x}} = 0$, which means $\phi_i^{\mathbf{x}}$ is perpendicular to the linear space spanned by $\{K_x : x \in \mathbf{x}\}$. In this case we do not have a representation of $\phi_i^{\mathbf{x}}$ with $\{K_x : x \in \mathbf{x}\}$. When $\lambda_i^{\mathbf{x}} > 0$, from $S_{\mathbf{x}}^T U_i = \frac{1}{\lambda_i^{\mathbf{x}}} S_{\mathbf{x}}^T (\frac{1}{m}\mathbb{K} U_i) = \frac{1}{\lambda_i^{\mathbf{x}}} L_K^{\mathbf{x}} (S_{\mathbf{x}}^T U_i)$, and $\left\| S_{\mathbf{x}}^T U_i \right\|_K^2 = m \left\langle U_i, \frac{1}{m}\mathbb{K} U_i \right\rangle_{\mathbb{R}^m} = m\lambda_i^{\mathbf{x}}$, we can take

$$\phi_i^{\mathbf{x}} = \frac{1}{\sqrt{m\lambda_i^{\mathbf{x}}}} S_{\mathbf{x}}^T U_i, \qquad U_i = \frac{1}{\sqrt{m\lambda_i^{\mathbf{x}}}} S_{\mathbf{x}} \phi_i^{\mathbf{x}}.$$

For two samples $\mathbf{x}$ and $\mathbf{u}$ with sizes m and n respectively, denote $\mathbb{K}_{\mathbf{u},\mathbf{x}}$ the $n \times m$ matrix of which the (i, j) entry is $K(u_i, x_j)$. Then $\mathbb{K}_{\mathbf{u},\mathbf{x}} = \mathbb{K}_{\mathbf{x},\mathbf{u}}^T$, and $S_{\mathbf{u}} S_{\mathbf{x}}^T = \mathbb{K}_{\mathbf{u},\mathbf{x}}$.

The Gram matrix $\mathbb{K}_{\mathbf{u},\mathbf{u}}$ of size $n \times n$ is similarly defined with the sample $\mathbf{u}$. The summary statistic $\mathbf{d}$ could be computed through

$$\begin{aligned} d_i &= \left\langle \phi_i^{\mathbf{u}}, \frac{1}{m}\sum_{j=1}^{m} y_j K_{x_j} \right\rangle_K = \left\langle \frac{1}{\sqrt{n\lambda_i^{\mathbf{u}}}} S_{\mathbf{u}}^T V_i, \frac{1}{m} S_{\mathbf{x}}^T \mathbf{y} \right\rangle_K \\ &= \frac{1}{m\sqrt{n\lambda_i^{\mathbf{u}}}} \left\langle V_i, \mathbb{K}_{\mathbf{u},\mathbf{x}}\mathbf{y} \right\rangle_{\mathbb{R}^n}, \end{aligned}$$

where $V = [V_1, \ldots, V_n]$ is the orthogonal matrix defined by the eigen-decomposition $\frac{1}{n}\mathbb{K}_{\mathbf{u},\mathbf{u}} = V\text{diag}\{\lambda_1^{\mathbf{u}}, \ldots, \lambda_n^{\mathbf{u}}\}V^T$.

1.3.2 Motivating Applications

Our work is inspired by two recent works [105, 66] in statistics. Consider the linear regression model $y = \mathbb{X}\beta + \varepsilon$, and its least squares solution $\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T y$. Roughly speaking, the basic idea in [105, 66] is only to collect the summary statistic $\mathbb{X}^T y$ as a whole, and use a new estimator $\hat{\beta}' = (\tilde{\mathbb{X}}^T\tilde{\mathbb{X}})^{-1}\mathbb{X}^T y$ to replace $\hat{\beta}$. Here $\tilde{\mathbb{X}}$ is the coefficient matrix made by openly accessible and unlabeled data without privacy issues. Real applications with data of both $\mathbb{X}$ and $\tilde{\mathbb{X}}$ are studied in the works. The relation between the predicted function $f_\lambda^{\mathbf{z}}$ of regularized least squares, and the predicted function $f_\lambda^{\mathbf{u},\mathbf{z}}$ of LESS is similar to that between $\hat{\beta}$ and $\hat{\beta}'$. In fact, for any $f, g, h \in \mathcal{H}_K$, define $f \otimes g$ as an operator by $(f \otimes g)h = \langle g, h \rangle_K f$. Define $P_N : \mathcal{H}_K \to \mathcal{H}_K$ as the orthogonal projection onto the subspace spanned by $\{\phi_i^{\mathbf{u}}\}_{i=1}^N$. That is, $P_N = \sum_{i=1}^N \phi_i^{\mathbf{u}} \otimes \phi_i^{\mathbf{u}}$. It is well known [90] that $f_\lambda^{\mathbf{z}} = (L_K^{\mathbf{x}} + \lambda I)^{-1} \frac{1}{m} S_{\mathbf{x}}^T \mathbf{y}$, and we can write $f_\lambda^{\mathbf{u},\mathbf{z}}$ by replacing $L_K^{\mathbf{x}}$ by $L_K^{\mathbf{u}}$, and inserting the projection P_N as a protocol,

$$f_\lambda^{\mathbf{u},\mathbf{z}} = (L_K^{\mathbf{u}} + \lambda I)^{-1} P_N \frac{1}{m} S_{\mathbf{x}}^T \mathbf{y}.$$

LESS can be used as a privacy-friendly substitute for regularized least squares (1.1). The solution $f_\lambda^{\mathbf{z}}$ in (1.2) of Problem (1.1) is a linear combination of kernel functions on the sample. To compute $f_\lambda^{\mathbf{z}}$, the sample $\mathbf{z}$ must be collected from the holder of data. To ship $f_\lambda^{\mathbf{z}}$ to the users, at least the input part $\mathbf{x}$ should explicitly be shipped, and the labels y_i's could thus be estimated via $y_i \approx f_\lambda^{\mathbf{z}}(x_i)$. Although when the input space X is an Euclidean domain with low dimension, one may ship $f_\lambda^{\mathbf{z}}$ in terms of its local approximations with splines or wavelets, such approximation could be difficult when the dimension of X is high. LESS solves this problem by collecting only the summary statistic $\mathbf{d}$ and shipping the predicted function $f_\lambda^{\mathbf{u},\mathbf{z}}$ in terms of the linear combination of $\phi_i^{\mathbf{u}}$'s, which is eventually the linear combination of K_{u_i}'s, with $u_i \in \mathbf{u}$ free of privacy issues.

The dimension N of the summary statistic $\mathbf{d}$ balances the protection of privacy, and the least squares error of the predicted function $f_\lambda^{\mathbf{u},\mathbf{z}}$. As suggested by Assumption (A4) and Corollary 1.1, if N is large enough such that $\lambda_{N+1} \leqslant \kappa^2\lambda$, $\mathbf{d}$ contains sufficient information that supports the optimal learning rate. In many applications the eigenvalues of L_K decay quickly and we do not need a large N to achieve (A4). For example, if X is an Euclidean domain and K is Sobolev smooth, then λ_i's decay polynomially [78]. If K is analytic, such as the widely used Gaussian kernel, then λ_i's decay exponentially [65]. From the proof of Theorem 1.1 and Corollary 1.1, we see that empirically, Assumption (A4) can be replaced by $\lambda_{N+1}^{\mathbf{u}} \leqslant \kappa^2\lambda$ without affecting the error estimate. A better privacy protection can be achieved by adding noise to $\mathbf{d}$ (or to $\mathbf{d}^j$'s under the distributed setting). We leave the quantitative analysis of this approach as future work.

For the case the sample $\mathbf{z}$ is held separately by ℓ different sources $\mathbf{z} = \cup_{i=1}^{\ell}\mathbf{z}_i$, there are recent works [19, 64, 45] that study the method of inflating each sub-sample $\mathbf{z}_i$ with a separate unlabeled sample. The inflation is done as follows. Suppose $\mathbf{u}$ is an unlabeled sample divided into ℓ subsets $\mathbf{u} = \cup_{i=1}^{\ell}\mathbf{u}_i$. For each i, all the sample points in $\mathbf{u}_i$ are equipped with a fake label 0, and all the labels in $\mathbf{z}_i$ are scaled by the factor $(|\mathbf{z}_i| + |\mathbf{u}_i|)/|\mathbf{z}_i|$ to compensate for these fake labels. Then $\mathbf{z}_i$ and $\mathbf{u}_i$ are mixed as a sample to yield an output function $f_\lambda^{\mathbf{u}_i \cup \mathbf{z}_i}$ from regularized least squares. The overall output function $\bar{f}_\lambda^{\mathbf{z}}$ is the weighted average of $f_\lambda^{\mathbf{u}_i \cup \mathbf{z}_i}$'s. By this operation, [64] proved (with the assumptions $|\mathbf{z}_1| = \ldots = |\mathbf{z}_\ell|$ and $|\mathbf{u}_1| = \ldots = |\mathbf{u}_\ell|$) that when

$$\ell \leqslant \frac{1}{\log^5 m + 1} \min \left\{ (n+m)^{1/2} m^{-\frac{s+1}{4r+2s}}, (n+m)^{1/3} m^{\frac{2r+s-2}{6r+3s}} \right\}, \tag{1.12}$$

the output function $\bar{f}_\lambda^{\mathbf{z}}$ still achieves the minimax optimal learning rate.

Compared with the inflation method studied in [19, 64, 45], LESS provides a way better solution to the learning problems with multiple sources of data. First, although for the scenarios where it is not allowed to bring together the training data from different sources, the distributed-learning setting solves the training problem, one still has to ship out the new instances (to different sources of training data) for prediction. Usually, these instances also contain private information, and it is not appropriate to circulate them around. Second, in the worst case scenario, when the sample size of each subset $\mathbf{z}_i$ is $O(1)$, and without loss of generality we use $\ell = m$, then (1.12) implies (recall $0 < s \leqslant 1$)

$$n \gtrsim m^{2+\frac{2}{2r+s}}, \tag{1.13}$$

where $n \gtrsim f(m)$ means there exists some positive constant $0 < C < \infty$ such that $n = n(m) \geqslant Cf(m)$ for any positive integer m. Note that in Corollary 1.1, the functional relation $n(m)$ is implicitly given by the lower bound $n \geqslant \max\{m, m^{\frac{2}{2r+s}}\}$. The restriction (1.13) requires much more unlabeled sample points than LESS does

$$n \geqslant \max\{m, m^{\frac{2}{2r+s}}\}, \tag{1.14}$$

in Corollary 1.1. Third, when (1.13) is satisfied, in each single computing node (located at the corresponding data source), according to the analysis in [64], the regularized least squares algorithm would process an inflated sample of size

$$\frac{n}{m} \gtrsim m^{1+\frac{2}{2r+s}}. \tag{1.15}$$

While for LESS, since the computation is centralized, we do not need significant computation provided by the data sources, and the sample size to be processed by the central computing node for LESS could be reduced, as suggested by (1.14), to

$$O\left(\max\left\{m, m^{\frac{2}{2r+s}}\right\}\right),$$

which is even much smaller than (1.15).

Chaudhuri et al. [20] studied an algorithm that uses random features (instead of the empirical features we use) for learning. Noise is added to the coefficients of the random features to achieve differential privacy. Because of the adoption of random features, this algorithm in [20] works only with translation invariant kernels.

1.4 Proof of the Main Theorem

We cite the following lemma from [11, Lemma E.4] and [9, Theorem IX.2.1].

Lemma 1.2. *Let A and B be positive definite operators on a separable Hilbert space $\mathcal{H}$. Write $\|\cdot\|_{\mathsf{op}(\mathcal{H})}$ the operator norm of $\mathcal{H}$. Then for any $0 \leqslant s \leqslant 1$, we have*

$$\|A^s B^s\|_{\mathsf{op}(\mathcal{H})} \leqslant \|AB\|_{\mathsf{op}(\mathcal{H})}^s. \tag{1.16}$$

Write $f_\lambda = (L_K + \lambda I)^{-1} L_K f_\rho$. One has $\lambda f_\lambda = L_K(f_\rho - f_\lambda)$. Write $\|\cdot\|_{\mathsf{op}}$ the operator norm of all the bounded linear operators on $\mathcal{H}_K$.

Lemma 1.3. *We have the following error bound*

$$\|f_\lambda^{\mathbf{u},\mathbf{z}} - P_N f_\lambda\|_\rho \leqslant \Omega_{\mathbf{u},\lambda}\left(R_\lambda^{\mathbf{z}} + \|f_\rho\|_K W_\lambda^{\mathbf{x}} + \|g_\rho\|_\rho W_\lambda^{\mathbf{u}}\right), \tag{1.17}$$

where

$$\Omega_{\mathbf{u},\lambda} := \left\|(L_K^{\mathbf{u}} + \lambda I)^{-1}(L_K + \lambda I)\right\|_{\mathsf{op}}, \tag{1.18}$$

$$R_\lambda^{\mathbf{z}} := \left\|(L_K + \lambda I)^{-1/2}\left(\frac{1}{m} S_{\mathbf{x}}^T \mathbf{y} - L_K^{\mathbf{x}} f_\rho\right)\right\|_K, \tag{1.19}$$

$$W_\lambda^{\mathbf{u}} := \left\|(L_K + \lambda I)^{-1/2}(L_K - L_K^{\mathbf{u}})\right\|_{\mathsf{op}}, \tag{1.20}$$

and $W_\lambda^{\mathbf{x}}$ is defined in the same way as (1.20) by substituting $\mathbf{u}$ with $\mathbf{x}$.

Proof. Since $\operatorname{span}\{\phi_i^{\mathbf{u}}\}_{i=1}^N$ is an invariant subspace of $L_K^{\mathbf{u}}$, P_N and $L_K^{\mathbf{u}}$ commute. We have

$$\begin{aligned}
& \left\|f_\lambda^{\mathbf{u},\mathbf{z}} - P_N f_\lambda\right\|_\rho \\
= & \left\|L_K^{1/2}(L_K^{\mathbf{u}} + \lambda I)^{-1} P_N \frac{1}{m} S_{\mathbf{x}}^T \mathbf{y} - L_K^{1/2} P_N f_\lambda\right\|_K \\
= & \left\|L_K^{1/2}(L_K^{\mathbf{u}} + \lambda I)^{-1/2} P_N (L_K^{\mathbf{u}} + \lambda I)^{-1/2} \left(\frac{1}{m} S_{\mathbf{x}}^T \mathbf{y} - (L_K^{\mathbf{u}} + \lambda I) f_\lambda\right)\right\|_K \\
= & \left\|L_K^{1/2}(L_K^{\mathbf{u}} + \lambda I)^{-1/2}\right\|_{\mathsf{op}} \left\|P_N\right\|_{\mathsf{op}} \left\|(L_K^{\mathbf{u}} + \lambda I)^{-1/2}(L_K + \lambda I)^{1/2}\right\|_{\mathsf{op}} \\
& \times \left\|(L_K + \lambda I)^{-1/2} \left(\frac{1}{m} S_{\mathbf{x}}^T \mathbf{y} - (L_K^{\mathbf{u}} + \lambda I) f_\lambda\right)\right\|_K . \qquad (1.21)
\end{aligned}$$

The right-hand side of (1.21) is the product of four norms. Below we bound them one by one. First, obviously $\|P_N\|_{\mathsf{op}} \leqslant 1$. Since $\lambda > 0$, for any $f \in \mathcal{H}_K$,

$$\langle f, L_K f\rangle_K \leqslant \langle f, (L_K + \lambda I) f\rangle_K .$$

Therefore we apply Lemma 1.2 to bound the first and the third factor of the right-hand side of (1.21) by $\Omega_{\mathbf{u},\lambda}^{1/2}$.

$$\begin{aligned}
\left\|L_K^{1/2}(L_K^{\mathbf{u}} + \lambda I)^{-1/2}\right\|_{\mathsf{op}} &= \left\|(L_K^{\mathbf{u}} + \lambda I)^{-1/2} L_K (L_K^{\mathbf{u}} + \lambda I)^{-1/2}\right\|_{\mathsf{op}}^{1/2} \\
&\leqslant \left\|(L_K^{\mathbf{u}} + \lambda I)^{-1/2} (L_K + \lambda I)(L_K^{\mathbf{u}} + \lambda I)^{-1/2}\right\|_{\mathsf{op}}^{1/2} \\
&= \left\|(L_K^{\mathbf{u}} + \lambda I)^{-1/2}(L_K + \lambda I)^{1/2}\right\|_{\mathsf{op}} \leqslant \Omega_{\mathbf{u},\lambda}^{1/2}. \qquad (1.22)
\end{aligned}$$

Since $r \geqslant 1/2$, we cite from [90] the bound that $\|f_\lambda\|_K \leqslant \|g_\rho\|_\rho$. Consider the following decomposition

$$\frac{1}{m} S_{\mathbf{x}}^T \mathbf{y} - (\lambda I + L_K^{\mathbf{u}}) f_\lambda = \left(\frac{1}{m} S_{\mathbf{x}}^T \mathbf{y} - L_K^{\mathbf{x}} f_\rho\right) + \left(L_K^{\mathbf{x}} - L_K\right) f_\rho + \left(L_K - L_K^{\mathbf{u}}\right) f_\lambda,$$

which leads to the bound $R_\lambda^{\mathbf{z}} + \|f_\rho\|_K W_\lambda^{\mathbf{x}} + \|g_\rho\|_\rho W_\lambda^{\mathbf{u}}$ of the fourth factor of the right-hand side of (1.21), and thus completes the proof. □

Lemma 1.4. *Let* $1/2 \leqslant r \leqslant 1$ *and* $\lambda > 0$*. We have*

$$\|P_N f_\lambda - f_\lambda\|_\rho \leqslant \Omega_{\mathbf{u},\lambda}^r (\lambda_{N+1}^{\mathbf{u}} + \lambda)^r \|g_\rho\|_\rho. \tag{1.23}$$

Proof. Recall that P_N and $L_K^{\mathbf{u}}$ commute. In particular,

$$\begin{aligned}(I - P_N)(L_K^{\mathbf{u}} + \lambda I)^r &= \left(\sum_{i \geqslant N+1} \phi_i^{\mathbf{u}} \otimes \phi_i^{\mathbf{u}}\right)\left(\sum_{j\geqslant 1} (\lambda_j^{\mathbf{u}} + \lambda)^r \phi_j^{\mathbf{u}} \otimes \phi_j^{\mathbf{u}}\right) \\ &= \sum_{j \geqslant N+1} (\lambda_j^{\mathbf{u}} + \lambda)^r \phi_j^{\mathbf{u}} \otimes \phi_j^{\mathbf{u}},\end{aligned}$$

so $\|(I - P_N)(L_K^{\mathbf{u}} + \lambda)^r\|_{\mathsf{op}} = (\lambda_{N+1}^{\mathbf{u}} + \lambda)^r$. By Lemma 1.2 and Inequality (1.22), we have

$$\begin{aligned}&\|P_N f_\lambda - f_\lambda\|_\rho \\ =\ & \left\|L_K^{1/2}(I - P_N) L_K^{\frac{1}{2}+r}(L_K + \lambda I)^{-1} L_K^{1/2} g_\rho\right\|_K \\ =\ & \left\|L_K^{1/2}(L_K^{\mathbf{u}} + \lambda I)^{-1/2}\right\|_{\mathsf{op}} \left\|(L_K^{\mathbf{u}} + \lambda I)^{1/2}(I - P_N)(L_K^{\mathbf{u}} + \lambda I)^{r-\frac{1}{2}}\right\|_{\mathsf{op}} \\ &\times \left\|(L_K^{\mathbf{u}} + \lambda I)^{-(r-\frac{1}{2})}(L_K + \lambda I)^{r-\frac{1}{2}}\right\|_{\mathsf{op}} \left\|L_K^{r+\frac{1}{2}}(L_K + \lambda I)^{-(r+\frac{1}{2})}\right\|_{\mathsf{op}} \|g_\rho\|_\rho \\ \leqslant\ & \Omega_{\mathbf{u},\lambda}^r (\lambda_{N+1}^{\mathbf{u}} + \lambda)^r \|g_\rho\|_\rho.\end{aligned}$$

The proof is complete. □

The following lemma is from [47, Proposition 1]. It is a powerful tool recently developed [63, 47] for the analysis of kernel-based regularized least squares and related algorithms.

Lemma 1.5. *Let* $\lambda > 0$ *and* $0 < \delta < 1$*. One has with confidence at least* $1 - \delta$ *that*

$$\Omega_{\mathbf{u},\lambda} \leqslant \frac{2}{\lambda} \mathcal{B}_{n,\lambda}^2 \log^2 \frac{2}{\delta} + 2. \tag{1.24}$$

Denote $\mathsf{HS}(\mathcal{H}_K)$ the Hilbert space of all the Hilbert-Schmidt operators on $\mathcal{H}_K$. Write $\|\cdot\|_{\mathsf{HS}}$ the norm of $\mathsf{HS}(\mathcal{H}_K)$. In the following lemma, Item 1 is the well-known Hoffman-Wielandt inequality [51, 55, 10], and Item 2 is a standard corollary of Pinelis' vector-valued concentration inequality [75]. Detailed proof of Item 2 is available in [100, Proposition 5.3]. See also [53, 7, 17, 63, 90, 100, 110].

Lemma 1.6. *1. We have*

$$\sum_{i=1}^{\infty}(\lambda_i - \lambda_i^{\mathbf{x}})^2 \leqslant \|L_K - L_K^{\mathbf{x}}\|_{\mathsf{HS}}^2. \tag{1.25}$$

2. For $0 < \delta < 1$, we have with confidence at least $1-\delta$ that

$$\|L_K - L_K^{\mathbf{x}}\|_{\mathsf{HS}} \leqslant \frac{4\kappa^2}{\sqrt{m}} \log\frac{2}{\delta}. \tag{1.26}$$

For the following Lemma 1.7, the proof of (1.27) is available in [17]. The proof of (1.28) is available in [63, Lemma 17]. The bound (1.29) follows directly from Lemma 1.6 by substituting $\mathbf{x}$ with $\mathbf{u}$, and $m = |\mathbf{x}|$ with $n = |\mathbf{u}|$.

Lemma 1.7. *Let $0 < \delta < 1$. Each of the following bounds holds with confidence at least $1-\delta$.*

$$R_\lambda^{\mathbf{z}} \leqslant \frac{M+\sigma}{\kappa}\mathcal{B}_{m,\lambda}\log\frac{2}{\delta}, \tag{1.27}$$

$$W_\lambda^{\mathbf{u}} \leqslant \mathcal{B}_{n,\lambda}\log\frac{2}{\delta}, \quad \textit{and} \tag{1.28}$$

$$\lambda_i^{\mathbf{u}} \leqslant \lambda_i + \frac{4\kappa^2}{\sqrt{n}}\log\frac{2}{\delta}, \quad \textit{for all } i = 1, 2, \cdots. \tag{1.29}$$

Proof of Theorem 1.1. Recall that $1/2 \leqslant r \leqslant 1$. By Lemma 1.3 and Lemma 1.4,

$$\begin{aligned}
\left\|f_\lambda^{\mathbf{u},\mathbf{z}} - f_\rho\right\|_\rho &\leqslant \left\|f_\lambda^{\mathbf{u},\mathbf{z}} - P_N f_\lambda\right\|_\rho + \left\|P_N f_\lambda - f_\lambda\right\|_\rho + \left\|f_\lambda - f_\rho\right\|_\rho \\
&\leqslant \Omega_{\mathbf{u},\lambda}(R_\lambda^{\mathbf{z}} + \|f_\rho\|_K W_\lambda^{\mathbf{x}} + \|g_\rho\|_\rho W_\lambda^{\mathbf{u}}) \\
&\quad + \Omega_{\mathbf{u},\lambda}^r (\lambda_{N+1}^{\mathbf{u}} + \lambda)^r \|g_\rho\|_\rho + \lambda^r \|g_\rho\|_\rho,
\end{aligned} \tag{1.30}$$

where we have used the estimate $\|f_\lambda - f_\rho\|_\rho \leqslant \lambda^r \|g_\rho\|_\rho$ (see [90]). Let $0 < \delta < \frac{1}{5}$, then $\log \frac{2}{\delta} > \log 10 > 1$. From Lemma 1.5 and Lemma 1.7, we have with confidence at least $1 - \delta$ that (1.24), (1.27), (1.28) (for both $W_\lambda^{\mathbf{u}}$ and $W_\lambda^{\mathbf{x}}$ respectively), and (1.29) hold true simultaneously. Now we assume these five inequalities. Then

$$R_\lambda^{\mathbf{z}} + \|f_\rho\|_K W_\lambda^{\mathbf{x}} + \|g_\rho\|_\rho W_\lambda^{\mathbf{u}} \leqslant \left(\frac{M+\sigma}{\kappa} + \|f_\rho\|_K + \|g_\rho\|_\rho \right) \mathcal{B}_{m,\lambda} \log \frac{2}{\delta}.$$

We combine the argument above and (1.29) to continue the bound (1.30).

$$\begin{aligned} \|f_\lambda^{\mathbf{u},\mathbf{z}} - f_\rho\|_\rho \leqslant & \left(\frac{2\mathcal{B}_{n,\lambda}^2}{\lambda} + 2 \right) \left(\frac{M+\sigma}{\kappa} + \|f_\rho\|_K + \|g_\rho\|_\rho \right) \mathcal{B}_{m,\lambda} \log^3 \frac{2}{\delta} \\ & + \left(\frac{2\mathcal{B}_{n,\lambda}^2}{\lambda} + 2 \right)^r \left(\lambda + \frac{4\kappa^2}{\sqrt{n}} + \lambda_{N+1} \right)^r \|g_\rho\|_\rho \log^{3r} \frac{2}{\delta} + \|g_\rho\|_\rho \lambda^r, \end{aligned}$$

The proof is completed by scaling δ to $\delta/5$. □

Proof of Corollary 1.1. Recall that $1/2 \leqslant r \leqslant 1$, $n \geqslant m$, and $0 < s \leqslant 1$. With the assumption $\mathcal{N}(\lambda) \leqslant C_1 \lambda^{-s}$ and the setting $\lambda = m^{-\frac{1}{2r+s}}$, (1.8) implies

$$\mathcal{B}_{n,\lambda} \leqslant \mathcal{B}_{m,\lambda} \leqslant \frac{2\kappa^2}{m} m^{\frac{1/2}{2r+s}} + 2\kappa \sqrt{\frac{C_1}{m} m^{\frac{s}{2r+s}}} \leqslant 2\kappa(\kappa + \sqrt{C_1}) m^{-\frac{r}{2r+s}}, \tag{1.31}$$

so

$$\frac{\mathcal{B}_{n,\lambda}^2}{\lambda} \leqslant 4\kappa^2 (\kappa + \sqrt{C_1})^2 m^{-\frac{2r-1}{2r+s}} \leqslant 4\kappa^2 (\kappa + \sqrt{C_1})^2. \tag{1.32}$$

Recall the assumptions $\lambda_{N+1} \leqslant \kappa^2 \lambda$ and $n \geqslant m^{\frac{2}{2r+s}}$. Therefore $\frac{1}{\sqrt{n}} \leqslant m^{-\frac{1}{2r+s}} = \lambda$ and

$$\frac{4\kappa^2}{\sqrt{n}} + \lambda_{N+1} \leqslant 5\kappa^2 \lambda.$$

So, Theorem 1.1 implies that

$$\begin{aligned}
\|f_\lambda^{\mathbf{u},\mathbf{z}} - f_\rho\|_\rho &\leqslant \left(\frac{2\mathcal{B}_{n,\lambda}^2}{\lambda}+2\right)\left(\frac{M+\sigma}{\kappa}+\|f_\rho\|_K+\|g_\rho\|_\rho\right)\mathcal{B}_{m,\lambda}\log^3\frac{10}{\delta} \\
&\quad + \left(\frac{2\mathcal{B}_{n,\lambda}^2}{\lambda}+2\right)^r\left(\lambda+\frac{4\kappa^2}{\sqrt{n}}+\lambda_{N+1}\right)^r\|g_\rho\|_\rho\log^{3r}\frac{10}{\delta}+\|g_\rho\|_\rho\lambda^r \\
&\leqslant (8\kappa^2(\kappa+\sqrt{C_1})^2+2)(2\kappa(\kappa+\sqrt{C_1})) \\
&\quad \times\left(\frac{M+\sigma}{\kappa}+\|f_\rho\|_K+\|g_\rho\|_\rho\right)m^{-\frac{r}{2r+s}}\log^3\frac{10}{\delta} \\
&\quad +(8\kappa^2(\kappa+\sqrt{C_1})^2+2)^r(1+5\kappa^2)^r\|g_\rho\|_\rho m^{-\frac{r}{2r+s}}\log^{3r}\frac{10}{\delta} \\
&\quad +\|g_\rho\|_\rho m^{-\frac{r}{2r+s}} \\
&\leqslant C_2 m^{-\frac{r}{2r+s}}\log^3\frac{10}{\delta},
\end{aligned}$$

where $C_2 = (8\kappa^2(\kappa+\sqrt{C_1})^2+2)(2\kappa(\kappa+\sqrt{C_1}))\left(\frac{M+\sigma}{\kappa}+\|f_\rho\|_K+\|g_\rho\|_\rho\right)+(8\kappa^2(\kappa+\sqrt{C_1})^2+2)^r(1+5\kappa^2)^r\|g_\rho\|_\rho+\|g_\rho\|_\rho$. □

We would like to acknowledge Professor Jian Huang for the helpful discussions, in particular, the introduction of the works [105, 66] to us.

Chapter 2

Search for K: Assessing Five Topic-modeling Approaches to 120,000 Canadian Articles

2.1 Introduction

The past two decades have witnessed an explosion in methods, algorithms and tools designed to identify discussion topics in automated text analysis. Noteworthy among these research efforts, the Latent-Dirichlet-allocation (LDA) approach assuming a Dirichlet prior distribution assigns a specific set of topics to each document, based on a fixed number (K) of topics. By incorporating both observed and latent variables, this Bayesian generative method allows for latent processes to capture similarities among sets of observations and thus results in a more precise assignment of topics to documents (and words to documents) [14]. While this method has been further developed to detect the number of optimal discussion topics based on a nonparametric Bayesian model [94], in practice the ultimate decision on the choice of K still relies on significant input from domain experts. In a more recent review of data analysis with latent models, Blei highlights a tension between orthodox Bayesian thinking and model criticism [13]. While the former attempts to integrate all possible sources of uncertainties in a more complex mixture or "super" models, the latter tries to tell whether the essence of the data has been captured by model specification and/or parameter inference. Yet, model criticism is becoming increasingly challenging with the proliferation of latent models in that we do not necessarily know whether the data, model specification, or inference algorithms plays a more significant part in shaping the (approximate) posterior. In response to these issues, this research uses various topic-modeling approaches to assess the choice of K via different training methods, where model specification and inference algorithms play different roles in shaping research findings.

2.2 Preprocessing Techniques

2.2.1 Data Cleaning and Stopwords Removal

Before applying topic models, the corpus needs to be cleaned. We first removed the common stopwords in English [68] such as `the`, `a`, and `an`, then we apply RAKE [82] to combine words into phrases such that words like `united states` are combined as `united-states`.

2.2.2 Term Frequency-inverse Document Frequencies

To apply topic-modeling methods, we represent a large corpus of text using a document-word matrix X, where each column corresponds to a document and each row corresponds to a word [59]. Since a word's frequency in a corresponding document cannot suggest the word's relative importance in the whole corpus, elements of the document-word matrix are often weighted by term frequency-inverse document frequencies (tf-idf) [80]. One way to calculate the tf-idf weight $w_{t,d}$ associated with a term (word) t and a document d is as follows [2],

$$w_{t,d} = \mathsf{tf}_{t,d} \times \log \frac{N}{\mathsf{df}_t}$$

where $\mathsf{tf}_{t,d}$ is a term t's frequency in the document d, N is the total number of documents, and df_t is the total number of documents containing the term t. Clearly, $w_{t,d}$ increases if a term has a higher frequency in a document but such increase is offset by the term's prevalence across all documents in text corpus. This tf-idf weight thus tends to filter out common words or stopwords which appear to be popular in most documents.

2.3 Five Approaches to Topic Modeling

To guide our assessment of different approaches to topic modeling, we next briefly discuss methodological details of the five models being adopted in this research.

2.3.1 Latent Semantic Analysis

Theoretical Review

Based on singular value decomposition of the document-word matrix, latent semantic analysis (LSA) has long been adopted by scholars from different disciplines to identify topics and themes contained in text corpus [24]. This is achieved by providing a low-rank approximation to the previously defined word-document matrix X [39]. To understand how LSA works, we have its singular value decomposition (SVD) of X as:

$$X = U\Sigma V^T,$$

where both U and V are orthogonal matrices and Σ is a diagonal matrix. To further explore these three matrices, we first note that the square matrix XX^T contains all dot products denoting the correlation between any two word vectors across all documents, and X^TX contains all dot products denoting the correlation between any two document vectors. And we have:

$$U^T X X^T U = \Sigma\Sigma^T \text{and } V^T X^T X V = \Sigma^T\Sigma, \text{or}$$

$$X X^T = U\Sigma\Sigma^T U^T \text{and } X^T X = V\Sigma^T\Sigma V^T.$$

In other words, XX^T and X^TX have the same non-zero eigenvalues expressed by $\Sigma\Sigma^T$ (or, equally by $\Sigma^T\Sigma$), and their eigenvectors are contained in U and V, respectively.

Application in Topic Modeling

The number of positive singular values in Σ suggests the rank of X, or the number of topics in the current research setting, while the values of these singular values suggests the relative importance of these topics. For a space spanned by singular vectors corresponding to these singular values (i.e., topics), the coordinates of a word i across all topics are denoted by the i^{th} row of U and the coordinates of a document j across all topics are denoted by the j^{th} column of V^T. The corresponding loadings of all words on the k^{th} topic are given by elements in the k^{th} columns of U; and the corresponding loadings of all documents on the k^{th} topic are given by elements in the k^{th} rows of V^T. While topics identified by LSA can be viewed as clusters of words and/or documents once they are projected to a "semantic space", we use columns of U to denote topics (and their corresponding relations with words). If the values of singular values are small or below a certain threshold specified by researchers, it is possible to remove these singular values and achieve a low-rank approximation [93].

2.3.2 Principal Component Analysis

Theoretical Review

The idea of principal component analysis (PCA) is very similar to that of SVD [54]. For the document-word matrix X, PCA tries to project the data to orthogonal directions so that distinctive features from the data can be retained as much as possible. In other words, if the covariance matrix associated with X is given by XX^T, PCA is looking for a projection matrix P such that after the projection the covariance matrix Y^TY of the resulted new document-word matrix $Y = PX$ has the largest variance in these projection directions. Yet, one constraint in the search for P is that these projection directions suggested by P should be basis vectors and orthogonal to each other. Otherwise, the direction associated with the second largest variance will be always parallel to or even overlap with that associated with the largest variance (and so forth for the remaining directions), which provides little information of the data. As a consequence, the off-diagonal elements (i.e., covariance) of Y^TY should be zero and PCA essentially deals with an issue of optimization with a constraint. We have:

$$Y^TY = (PX)(PX)^T = PXX^TP^T = D$$

where D should be a diagonal matrix. Related to our discussion on SVD, if we rank eigenvectors $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n$ of XX^T and form a new matrix $Z = (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n)$ and let:

$$Z^TXX^TZ = \Sigma^T\Sigma = \Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} \tag{2.1}$$

D will be a diagonal matrix if we make $P = Z^T$. Therefore, the matrix containing all the eigenvectors of XX^T provides the loadings of all words on any topic and a solution to the application of PCA to topic modeling. The optimization issue also corresponds to the maximization of $\mathbf{z}_i^TXX^T\mathbf{z}_i$ when $\mathbf{z}_i^T\mathbf{z}_i = 1$. If we take the derivative of $\mathbf{z}_i^TXX^T\mathbf{z}_i - \lambda\mathbf{z}_i^T\mathbf{z}_i$ with respective to $\mathbf{z}_i$, we have $(XX^T - \lambda I)\mathbf{z}_i = 0$ and $\mathbf{z}_i$ must be an eigenvector of XX^T.

Application in Topic Modeling

We take these extracted principal components as topics, and extract the top words of these topics by finding the top corresponding values in the principal component.

To summarize, the relation between LSA and PCA is similar to that between maximum likelihood estimation and ordinary least squares estimation in linear regression settings: they appear to follow different principles yet (sometimes) yield the same result. Also, due to the fact that the components extracted by PCA or SVD are often mixed with positive and negative values, the interpretation of negative values can be less straightforward. Nevertheless, these two methods differ from each other in terms of computing: the calculation involving covariance matrices is demanding when observations and eigenvectors associated with PCA are large, while numerical methods can be readily applied to the calculation of SVD.

2.3.3 Factor Analysis

Theoretical Review

While PCA tries to identify major components embedded in the data matrix, factor analysis (FA) aims to represent the data matrix and its internal relations via latent factors (variables). To do so, FA draws on a parametric model and a series of assumptions/conditions. More specifically, if words in the document-word matrix X are centered on its means in a document and we obtain a new document-word matrix X_*, we try to express the p words using latent factors:

$$Y_{n\times p} = X_*^T = F_{n\times k}A_{k\times p} + \varepsilon_{n\times p}$$

where F is a matrix containing all (latent) factors $F_1, F_2, ..., F_k$ for each of n document, $A = (a_{ij})_{k\times p}$ is a loading matrix representing the loadings of all words on each of the k factors, and ε is the Gaussian error term. The FA model satisfies the following four assumptions/conditions:

1. The expectation and covariance (matrix) of F_i are 0 and I_n, respectively;
2. The expectation and covariance (matrix) of ε_i are 0 and $\sigma^2_{n\times n} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2)$;
3. The covariance between ε and F is 0;
4. $\mathrm{Cov}(Y_i) = AA^T + \sigma^2 I$ and $\mathrm{Cov}(Y_i, F_i) = A_{k\times p}$.

This conclusion that $\mathrm{Cov}(Y) = AA^T + \sigma^2 I$ has two implications. First, it is possible to calculate the loading matrix A first and then solve the latent factors using $F = \Sigma y A^T$. Second, for the i^{th} row a_i in A and a word y_i across all observations (i.e., documents), we have $\mathrm{var}(y_i) = a_i' a_i + \sigma_i^2$ and $\mathrm{Cov}(y_i, y_k) = a_i' a_k$. The sum of squared loadings of y_i on all factors, or $a_i' a_i$ (i.e., the common variance), denotes the dependence of y_i on all factors, or the extent to which y_i is explained by all factors.

Factor analysis can be implemented in different ways and this study adopts the EM algorithm to conduct factor analysis [42, 83]. Yet, in existing literature the link between PCA and FA has been particularly noted [24, 74]. Related to Equation (2.1), we have the eigenvalues of YY^T as $\lambda_1, \lambda_2, \cdots, \lambda_p$, their corresponding standardized eigenvectors as $\mathbf{z}_{y_1}, \mathbf{z}_{y_2}, \cdots, \mathbf{z}_{y_p}$, and $YY^T = \sum_{i=1}^p \lambda_i \mathbf{z}_{y_i} \mathbf{z}'_{y_i}$ given that:

$$\begin{aligned}
YY^T = \Lambda_Y &= Z_Y \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix} Z_Y^T \\
&= (\mathbf{z}_{y_1}, \mathbf{z}_{y_2}, \cdots, \mathbf{z}_{y_p}) \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix} \begin{pmatrix} \mathbf{z}'_{y_1} \\ \mathbf{z}'_{y_2} \\ \vdots \\ \mathbf{z}'_{y_p} \end{pmatrix} \\
&= (\sqrt{\lambda_1}\mathbf{z}_{y_1}, \sqrt{\lambda_2}\mathbf{z}_{y_2}, \cdots, \sqrt{\lambda_p}\mathbf{z}_{y_p}) \begin{pmatrix} \sqrt{\lambda_1}\mathbf{z}'_{y_1} \\ \sqrt{\lambda_2}\mathbf{z}'_{y_2} \\ \vdots \\ \sqrt{\lambda_p}\mathbf{z}'_{y_4} \end{pmatrix}
\end{aligned}$$

For the vector $(\sqrt{\lambda_1}\mathbf{z}_{y_1}, \sqrt{\lambda_2}\mathbf{z}_{y_2}, \cdots, \sqrt{\lambda_p}\mathbf{z}_{y_p})$, its first m entries (where $m < p$)

provides a possible solution to A and thus correspond to m latent factors because:

$$
\begin{aligned}
YY^T &\approx \hat{A}\hat{A}^T + \hat{\sigma}^2 \\
&= \lambda_1 \mathbf{z}_{y_1}\mathbf{z}'_{y_1} + \lambda_2 \mathbf{z}_{y_2}\mathbf{z}'_{y_2} + \cdots + \lambda_m \mathbf{z}_{y_m}\mathbf{z}'_{y_m} + \hat{\sigma}^2.
\end{aligned}
$$

Finally, it should be noted that these factors identified are often rotated to achieve maximum variance so that these independent factors can have better explanatory power.

Application in Topic Modeling

The obtained factors are considered as the weight vectors for each topic, we identify the top words according to the same principle as for SVD (LSA) and PCA, we sort the words according to its factor value and retain those with high values.

2.3.4 Non-negative Matrix Factorization

Theoretical Review

Non-negative matrix factorization (NMF) decomposes a matrix V into two matrices W and H and all elements of the three matrices are not negative [60]:

$$V_{n\times m} = W_{n\times r} H_{r\times m}$$

where the dimension of r is often much smaller than that of m and n. The NMF has a clear advantage over other similar algorithms in computing, interpretation and data storage. By making all elements in the three matrices non-negative, any column vector v_i in V can be expressed by a weighted sum of all column vectors in W and their corresponding weights are given by elements in the i^{th} column of H:

$$v_i = w_1 h_{1i} + w_2 h_{2i} + \cdots + w_r h_{ri} = W h_i.$$

In other words, we can learn how a whole system consists of different parts via these positive weights generated by NMF. The general idea behind NMF is also inherently related to how a whole system and its relations with different parts are perceived by human beings.

Application in Topic Modeling

The relation between NMF and topic modeling, especially probabilistic latent semantic analysis (PLSA), has been noted [40]. For the document-word matrix X, we could define elements of W as $w_{ik} = P(\text{topic}_k)P(\text{word}_i|\text{topic}_k)$, elements in H as $h_{kj} = P(\text{document}_j|\text{topic}_k)$ and have elements x_{ij} as:

$$\begin{aligned} x_{ij} &= \sum w_{ik}h_{kj} \\ &= \sum P(\text{topic}_k)P(\text{word}_i|\text{topic}_k)P(\text{document}_j|\text{topic}_k) \end{aligned}$$

The idea is similar to that of PLSA, where a probabilistic model is used to generate topics, and words/documents are further generated based on the topic distribution.

2.3.5 Latent Dirichlet Allocation (LDA)

Theoretical Review

In topic modeling, LDA provides a generative statistical model allowing for observed words and documents to be explained by latent topics that capture the similarities of words/documents [14]. For a text corpus, the generative process of LDA can be briefly summarized as follows. First, the (optimal) number of topics K needs to be specified. Second, a parameter θ_i which governs the distribution of K topics in the i^{th} document, is drawn from a Dirichlet prior distribution $D(\alpha)$. The hyper-parameter α is a K-dimensional vector with its elements (positive real numbers) denoting the relative weights of the K topics. Third, a parameter φ_k, which governs the distribution of all V words occurring in a topic k, is drawn from another Dirichlet prior distribution $D(\beta)$. The hyper-parameter β is a V-dimensional (sparse) vector with its elements denoting the relative weights of the V words. Finally, for a word

$X_{j,l}$ in the l^{th} location of the j^{th} document, its corresponding topic $t_{j,l}$ is drawn from a multinomial distribution $M(\theta_j)$ and the word is then generated from a multinomial distribution $M(\varphi_{t_{j,l}})$. The likelihood function of the model is:

$$\begin{aligned} & P(X, t, \theta, \varphi; \alpha, \beta) \\ & = \prod_{i=1}^{K} P(\varphi_i; \beta) \prod_{j=1}^{N} P(\theta_j; \alpha) \prod_{l=1}^{L_j} P(t_{j,l}|\theta_j) P(X_{j,l}|\varphi_{t_{j,l}}) \end{aligned}$$

We adopted the online variational Bayes algorithm [52] to optimize the model, by optimizing the Evidence Lower BOund (ELBO), details can be found in the reference.

Application in Topic Modeling

The LDA model is designed for topic modeling, therefore the connection is clear and simple: the estimated posterior φ represents the word distribution in each topic while the estimated posterior θ represents the topic distribution in each document.

2.4 Data and Measures

2.4.1 Data

The text corpus used in the current study was retrieved from three major newspapers in Canada with national influence: *The Globe and Mail, (The) Toronto Star and National Post*. All newspaper articles published in any of the three newspapers from January 1st 1977 to June 30th 2019 are retrieved as long as they contain the word "Chinese". The data retrieval process took place from 2017 to 2019. In total, 52,317, 43,529, and 23,634 articles were retrieved from *The Globe and Mail, Toronto Star and National Post*, respectively. Based on lists of stop words and results from preliminary data analysis, the research team performed multiple rounds of data cleaning and compiling to remove stop words and meaningless words for topic modeling (e.g., reporters' names, street address) prior to our analysis.

2.4.2 Measures

In search for the optimal number of topics K, we compare three types of measures to assess results estimated from the five topic-modeling methods: held-out likelihood (or reconstruction loss when applicable), coherence statistics, and graph-based dimensionality selection [18, 104, 73, 70].

Fitting Error Measure

We calculate the held-out likelihood of fitted models using 3-fold cross validation [5]. Specifically, we split the text corpus into three parts, treat one part as a test set and the other two as training sets. We repeat the estimation process for all three parts of the text corpus and calculate the average of the held-out likelihood/loss. We then compute either held-out likelihood or loss based on type of model, we have PCA, FA and LDA implemented as probabilistic models. It should be noted, however, the focus of the held-out-likelihood/loss approach is the predictive power of a specific model instead of the latent structure (e.g., topics) of the text corpus at stake. Also note that for log-likelihood, higher value indicates better performance, vice versa for reconstruction loss.

Coherence Statistics

Four measures of coherence are adopted in this study: C_v, C_{npmi}, C_{uci}, U_{mass} [81]. If a set of statements or terms mutually support each other, we say that this set of statements is coherent. For a specific topic, these coherence measures capture the degree of semantic similarity among words in the topic, thus allow scholars to assess whether topic modeling results represent actual semantic topics or statistical artifacts. We use the average of a coherence measure of each topic as a within-topic measure of topic coherence.

We first define the notion of pointwise mutual information:

$$PMI(x, y) = \log\left(\frac{P(x, y) + \epsilon}{P(x)P(y)}\right)$$

where ϵ is the smoothing constant and is often set to 1.

These four measures of coherence can be briefly described as follows. C_{uci} is probably the earliest statistic proposed to address topic coherence, which uses a sliding window and pointwise mutual information to measure the co-occurrence probability of every word pairs in a topic. It has been suggested that C_{uci} provides an extrinsic measure of coherence since it pairs every single word with every other word in the topic [73].

Suppose we have a topic of three words {a, b, c}. The co-occurrence probability of any two words would be calculated based on sliding windows, for example, if our text is "a is b", the virtual documents with a size 2 sliding window would be "a is","is b". In this case, $P(a) = \frac{1}{2}$ (appeared once in two virtual documents), $P(a, b) = 0$ (no co-occurrence of a and b), and

$$C_{uci} = \frac{1}{3}\left[PMI(a, b) + PMI(a, c) + PMI(b, c)\right]$$

C_{npmi} can be viewed as an enhanced version of C_{uci} because the former uses normalized pointwise mutual information (NPMI) instead of pointwise mutual information [3]. The NPMI is defined as the following:

$$NPMI(x, y) = \left(\frac{\log\frac{P(x,y)+\epsilon}{P(x)P(y)}}{\log(-P(x, y) + \epsilon)}\right)^{\gamma}$$

where ϵ is the smoothing constant and higher γ givers higher NPMI more weight.

C_v is proposed most recently and deals with indirect similarities between words [81], that is, some words should belong to the same topic but they rarely occur together; yet, their adjacent words should look similar. For example, suppose there are two statements “McDonald makes chicken nuggets” and “KFC serves chicken nuggets”, one will probably want to put McDonald and KFC together in the same topic. The mathematical details of C_v also appears to be somewhat complicated. The use of co-occurrence counts in the calculation of the NPMI of every top word to every other top word results in a set of vectors. For every top word, there is a corresponding vector. The indirect similarity is then calculated between the vector of every top word and the sum of all other top-word vectors. Cosine distance is used as a similarity measure.

Finally, based on the idea that the occurrence of every top word should be supported by every preceding top word, U_{mass} measures the conditional probability of weaker words given the presence of their corresponding stronger words in a topic. Different from the other three measures, U_{mass} is an intrinsic measure since the word list needs to be ordered and a word is compared only to its preceding and succeeding words [70]. To avoid the calculation of the logarithm of zero, a pairwise score function of the empirical conditional log-likelihood based on smoothing counts is used.

It it noteworthy that each coherence measure should be considered as independent therefore comparing intra-indicators is not meaningful. Also, all coherence indicators are higher the better.

Dimensionality Selection

The last measure originates from graph-based dimensionality selection. Since in methods like SVD (LSA) and PCA, we have a natural importance indicator which is the eigenvalue. People has used scree plots to identify the primary principal components, but given the very large dimensions (e.g., numbers of eigenvectors) associated with about 120,000 newspaper articles, the traditional threshold of dimensionality selection (eigenvalue as 1.0) cannot be readily applied to a big-data project. We thus relies on an automatic procedure, which maximizes a simple profile likelihood function, to search for the elbow point in a scree plot [104].

2.5 Results

The three types of measures based on results from the five methods of topic modeling are presented from Figure 2.1 to Figure 2.12. For the SVD (LSA) method, it is clear that the coherence statistics, especially for the C_{uci} and U_{mass} measures, favor fewer topics (see Figure 2.1). This opposite conclusion holds for the measure of held-out likelihood because more topics are associated with smaller errors (see Figure 2.3). Yet, according to the graph-based dimensionality selection, the optimal topics number appears to be 669 (see Figure 2.2).

Findings based on PCA are similar to these based on the SVD method. Coherence statistics, especially C_{uci} and U_{mass}, tend to suggest a smaller number of topics (see Figure 2.4). This pattern stands in contrast with the held-out likelihood, where the more the merrier (see Figure 2.6). The optimal number of topics suggested by dimensionality detection is 698 (see Figure 2.5). The coherence statistics for the FA method also prefer a smaller number of topics, although the value of U_{mass} slightly increases with a larger number of topics after 600 (see Figure 2.7). Yet, the held-out-likelihood measure of the FA model is able to specify the optimal number of topics, which appears to be 100 (see Figure 2.8).

The coherence statistics for the NMF methods reveal an interesting picture (see Figure 2.9). While the curves of C_{npmi} and C_v are relatively flat, results based on the C_{uci} and U_{mass} measures do not agree with each other: U_{mass} prefers a smaller number of topics but C_{uci} suggests that the value of K should be somewhere around

50 to 100. In Figure 2.10, the held-out error tends to support a larger number of optimal topics.

Finally, for the LDA method, the C_{npmi} and C_v measures do not show a strong preference over a particular number of topics (see Figure 2.11). The C_{uci} measure suggests that the value of K should be between 50 and 80 but the U_{mass} measure still favors a large number of topics. Finally, the held-out likelihood measure suggests that the optimal number of topics should be 20.

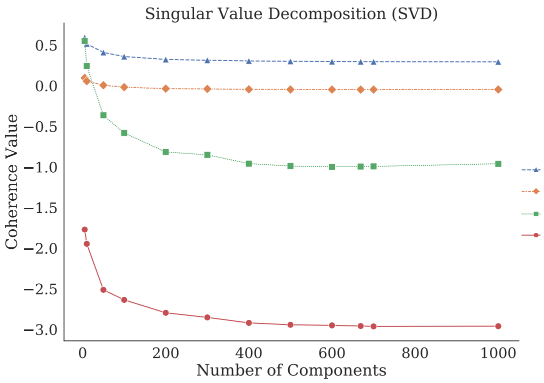

Figure 2.1: The SVD (LSA) method: Coherence.

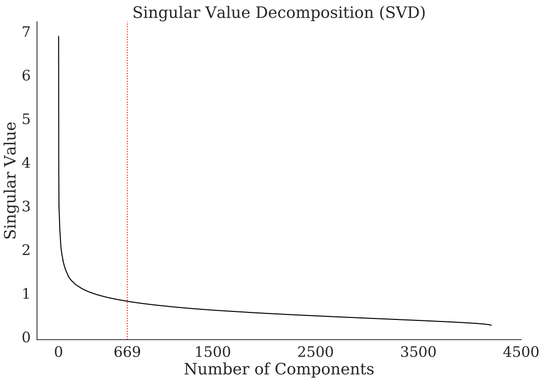

Figure 2.2: The SVD (LSA) method: Dimensionality selection.

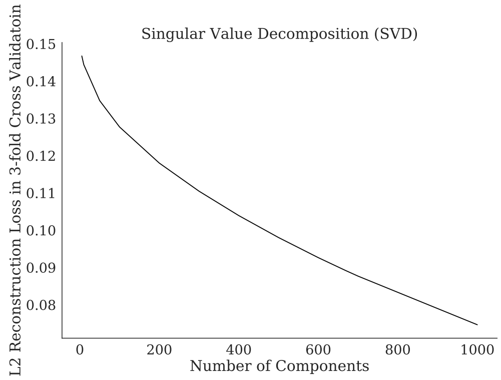

Figure 2.3: The SVD (LSA) method: Held-out error.

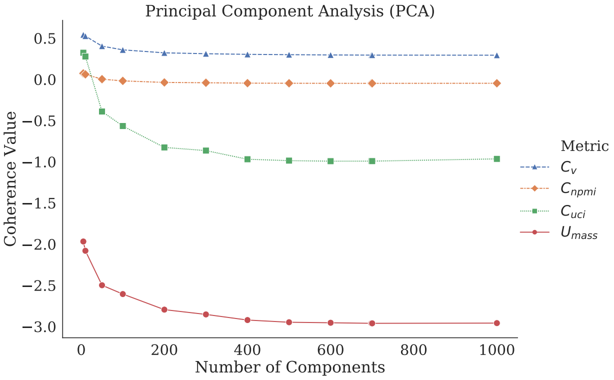

Figure 2.4: The PCA method: Coherence.

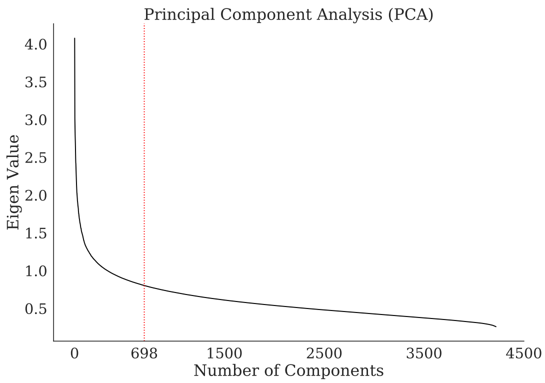

Figure 2.5: The PCA method: Dimensionality selection.

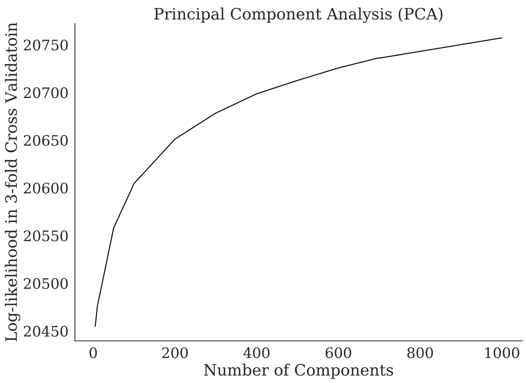

Figure 2.6: The PCA method: held-out likelihood.

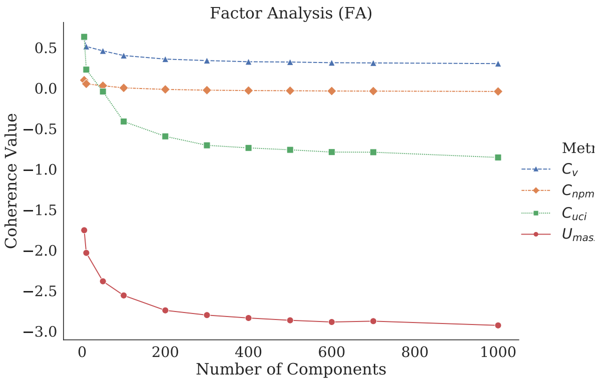

Figure 2.7: The FA method: Coherence.

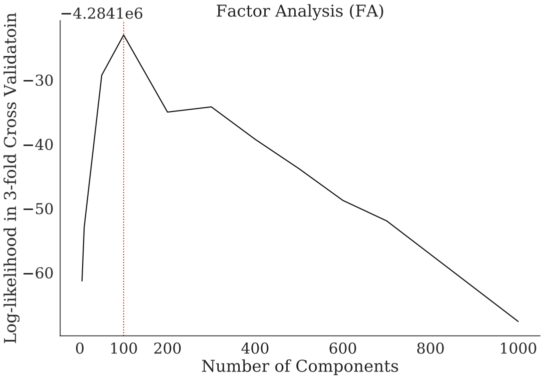

Figure 2.8: The FA method: held-out likelihood.

2.6 Conclusion

Based on an application of five approaches to topic modeling of about 120,000 newspaper articles in Canada, major findings comparing from three measures for the optimal number of topics can be summarized in Table 2.1. It should be noted, however, these findings are based on a specific text corpus and can vary if other forms of data are used.

As suggested by Table 2.1, when two approaches of topic modeling are methodologically similar to each other (i.e., SVD and PCA), these measures tend to report comparable results. Yet, the optimal number of topics can vary greatly across dif-

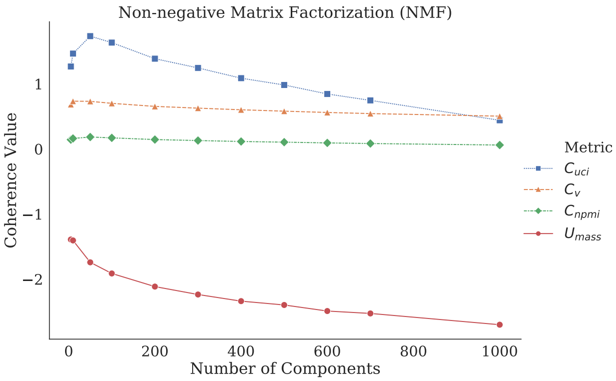

Figure 2.9: The NMF method: Coherence.

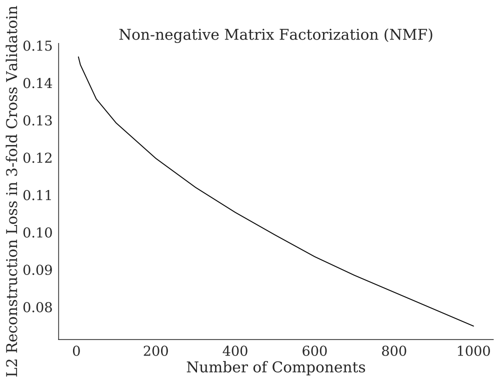

Figure 2.10: The NMF method: Held-out error.

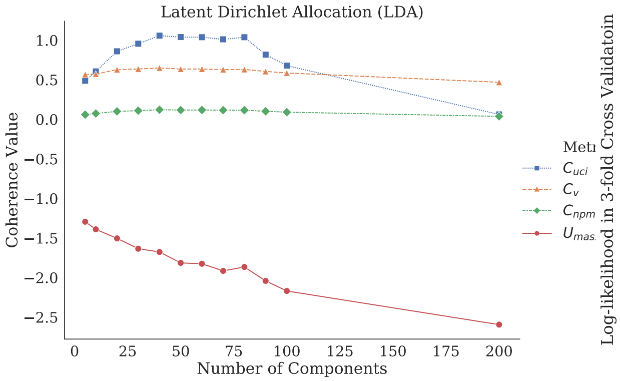

Figure 2.11: The LDA method: Coherence.

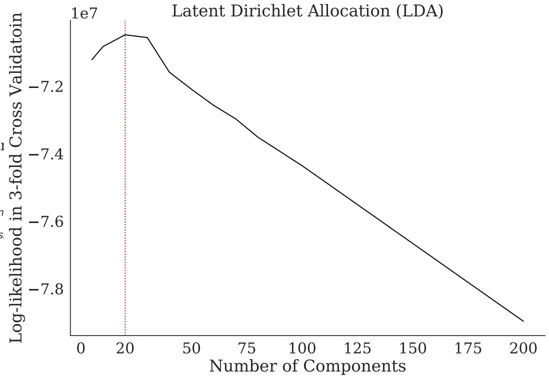

Figure 2.12: The LDA method: Held-out likelihood.

ferent approaches and measures. For the same method of topic modeling, different assessment measures can also suggest different and even opposite conclusions. Among these five topic modeling methods being investigated, only assessment measures pertaining to LDA modeling tend to suggest similar numbers of optimal topics. These interesting findings beg a key question in the search for an optimal number of topics: why should measures and methods based on different methodological philosophies and computing algorithms report similar, if not identical, numbers of optimal topics? Is there, in fact, an optimal number of topics to be discovered by more advanced methods? Without the input from domain experts, to what extent should optimal numbers of topics be viewed as methodological artifacts or distinctive features of the

Table 2.1: A summary of optimal number of topics suggested by different measures and methods

	SVD	PCA	FA	NMF	LDA
C_{uci}	Small	Small	Small	50+	50-80
C_v	Small*	Small*	Small*	50-	25*
C_{npmi}	Small*	Small*	Small*	50-	25*
U_{mass}	Small	Small	Small	Small	Small
Held-out likelihood (loss)	Large	Large	100	Large	20
Dimensionality selection	669	698	NA	NA	NA

Note: *possibly related to the scale of graphs, the conclusion suggested by this measure may not be very clear.

text corpus at stake? While the current study cannot answer all these questions, our mixed findings seem to suggest that *optimality* should be first defined in terms of, but not limited to, data reduction, latent structure, or predictive power, before any search for optimal topics takes place.

Chapter 3

Hurdle Model for Grouped and Right-censored Counts

3.1 Introduction

The modeling of count data has been an important field in applied mathematics, statistics, and social sciences [88, 1, 16, 101, 23, 43, 50]. Over the last few decades, several statistical models guided by different principles have been developed, implemented, tested, and applied by scholars to analyze count data across various fields of research, which include but not limited to Poisson models, negative binomial models, hurdle models, zero-inflated models [34, 31, 32, 15, 49, 56, 76, 84, 109, 4, 57, 98]. One major reason for an explosion of methods for modeling count data is that the observed distributions of counts are often dispersed and with excessive zeros. A concrete understanding of the source of over-dispersed count data is warranted to account for excessive zeros beyond that expected by a theoretical distribution [6, 77, 98].

When counts are treated as covariates (or independent variables) in empirical research, they can be easily analyzed as categorical variables by adopting an appropriate data coding method, such as dummy coding, effects coding, or spline regression [41, 72, 33, 62, 107, 108, 106]. Moreover statistical methods including ridge regression, principle component regression, cross-classified mixed-effect models, and the least absolute shrinkage and selection operator (the lasso) have been developed in the presence of collinearity among these categorical variables [58, 79, 36, 37, 38, 99, 27, 67, 69, 71, 95]. When counts are used as outcome variables, the modeling of counts becomes a complex issue especially with the presence of grouped and right-censored counts [28, 44, 76, 29].

In survey methodology, ordered selections with grouped and right-censored counts are often used to collect information on sensitive topics or from individuals with less cognitive capacities, such as children, the depressed, or the elderly [86, 87, 85]. Although ordered selections with grouped and right-censored counts are shown to be a valid and popular tool in data collection, the analysis of such data structure has been challenged by the absence of algorithms, programs, and models in analyzing grouped and right-censored counts. Although several Poisson-based methods have been recently proposed by pioneering studies to model grouped and right-censored counts in surveys [28, 76, 29], the hurdle model have not been specifically considered in the research context of grouped and right-censored counts. By adopting a truncated Poisson distribution, the hurdle model uses a different way to consider excessive zeros, which have not been fully considered in zero-inflated Poisson or negative binomial models. Moreover, as well demonstrated in existing literature, the hurdle model provides a flexible way to model counts with zero-inflation because both inflation and deflation of zeros can be considered [109, 16]. Next, we develop a general approach to consider hurdle models in the context of grouped and right-censored counts in surveys.

3.2 Methods

3.2.1 Hurdle Models for Count Data

In this work, count observations are those taking values from the set $\mathbb{N} := \{0, 1, \ldots\}$ of all the non-negative integers. For example, let Y be a random variable that has a Poisson distribution $Y \sim \text{Pois}(\mu)$ with mean $\mu > 0$, then

$$\text{Prob}(Y = k) = e^{-\mu}\frac{\mu^k}{k!}, \text{ for any } k \text{ in } \mathbb{N}.$$

One has $\mathbb{E}(Y) = \text{Var}(Y) = \mu$. With a sample $\{y_i\}_{i=1}^N$ drawn independently from $\text{Pois}(\mu)$, the method of moment estimator $\hat{\mu}_{\mathsf{MME}} = \bar{y} := \frac{1}{n}\sum_{i=1}^n y_i$ coincides with the maximum likelihood estimator (MLE) $\hat{\mu}_{\mathsf{MLE}} = \bar{y}$.

Another example is the negative binomial distribution $\text{NB}(\mu, \nu)$, where $\mu, \nu > 0$. Let $Y \sim \text{NB}(\mu, \nu)$, then

$$\text{Prob}(Y = k) = \frac{\Gamma(k + \nu)}{k!\Gamma(\nu)}\pi^\nu(1 - \pi)^k, \text{ for any } k \text{ in } \mathbb{N}, \tag{3.1}$$

where $\Gamma(x) = \int_0^\infty e^{-t}t^{x-1}dt$ for $x > 0$ is the gamma function, and $\pi = \frac{\nu}{\nu+\mu}$. One has $\mathbb{E}(Y) = \mu$ and $\text{Var}(Y) = \mu + \frac{\mu^2}{\nu}$. As $\nu \to \infty$, $\text{NB}(\mu, \nu)$ converges in law to $\text{Pois}(\mu)$. The MLE of μ is $\hat{\mu}_{\mathsf{MLE}} = \bar{y}$, but to our best knowledge, there is no closed-form expression of $\hat{\nu}_{\mathsf{MLE}}$. There are many ways of parameterization of negative binomial distributions in the literature. For example, (π, ν) in (3.1) is already a different parameterization. Also, we use $\alpha = \nu$ and $\beta = \mu/\nu$ to have [26]

$$\text{Prob}(Y = k) = \frac{\Gamma(\alpha + k)}{k!\Gamma(\alpha)}\frac{\beta^k}{(1 + \beta)^{k+\alpha}}, \text{ for any } k \text{ in } \mathbb{N}.$$

In real applications, it is usually observed from data that $\mathbb{E}(Y)$ and $\text{Var}(Y)$ are different. This suggests that Poisson is not the true model for data generation and sometimes $\text{NB}(\mu, \nu)$ is adopted. In some other scenarios, where the count at zero is disproportionately large or small, the hurdle Poisson model $\mathrm{HP}(\mu, p)$ is often employed. In particular, for $Y \sim \mathrm{HP}(\mu, p)$ with $\mu > 0$ and $0 < p < 1$, one has

$$\mathrm{Prob}(Y = k) = \begin{cases} 1 - p, & k = 0, \\ \frac{p}{1-e^{-\mu}} e^{-\mu} \frac{\mu^k}{k!}, & k \geqslant 1. \end{cases}$$

So, the hurdle model $\mathrm{HP}(\mu, p)$ assigns probability $1 - p$ to zero, and the conditional distribution for $k \geqslant 1$ is the truncated Poisson. We see that when $1 - p = e^{-\mu}$, $\mathrm{HP}(\mu, p) = \mathrm{Pois}(\mu)$. When $1 - p > e^{-\mu}$, we say that the zero outcome is inflated. When $1 - p < e^{-\mu}$, we say that the zero outcome is deflated.

Let $f(k)$ be a general probability mass function supported on $\mathbb{N}$. That is, f is defined on $\mathbb{N}$ with $f(k) > 0$ for all $k \in \mathbb{N}$, and $\sum_{k \in \mathbb{N}} f(k) = 1$. The corresponding hurdle model with a parameter $p \in (0, 1)$ is the distribution with the mass function f_p on $\mathbb{N}$, defined by

$$f_p(k) = \begin{cases} 1 - p, & k = 0, \\ \frac{p}{1-f(0)} f(k), & k \geqslant 1. \end{cases}$$

There is another model for the inflated zero outcome, called the zero inflated model, that parallels the hurdle model. In particular, the zero inflated model $f_{\mathsf{ZI},\tilde{p}}$ with parameter $0 < \tilde{p} < 1$, is a probability mass function on $\mathbb{N}$ defined by

$$f_{\mathsf{ZI},\tilde{p}} = \begin{cases} 1 - \tilde{p} + \tilde{p} f(0), & k = 0, \\ \tilde{p} f(k), & k \geqslant 1. \end{cases}$$

We see that when $p < 1 - f(0)$, $f_p = f_{\mathsf{ZI},\tilde{p}}$ with

$$\tilde{p} = \frac{p}{1 - f(0)}.$$

For the case $p \geqslant 1 - f(0)$, the hurdle model has no longer a $f_{\mathsf{ZI},\tilde{p}}$ representation. In particular, when $p = 1 - f(0)$, $f_p = f$. When $p > 1 - f(0)$, one has $1 - p < f(0)$, which corresponds to the distribution with deflated zero outcome. In this sense, the hurdle model provides a more flexible characterization of data. If the mean of the distribution f is μ, then the means μ_p and $\mu_{\mathsf{ZI},\tilde{p}}$ for the distributions f_p and $f_{\mathsf{ZI},\tilde{p}}$ are respectively,

$$\mu_p = \frac{p}{1-f(0)}\mu, \quad \text{and } \mu_{\mathsf{ZI},\tilde{p}} = \tilde{p}\mu.$$

Furthermore, if the variance of the distribution f is σ^2, then the variances of the distributions f_p and $f_{\mathsf{ZI},\tilde{p}}$ are respectively,

$$\sigma_p^2 = \frac{p\sigma^2}{1-f(0)} + \left[\frac{p}{1-f(0)} - \frac{p^2}{(1-f(0))^2}\right]\mu^2, \quad \text{and}$$
$$\sigma^2_{\mathsf{ZI},\tilde{p}} = \tilde{p}\sigma^2 + (\tilde{p}-\tilde{p}^2)\mu^2.$$

3.2.2 Regression Analysis of Hurdle Models with Grouped and Right-censored Data

We consider a family $\{f(k;\boldsymbol{\theta})\}$ of distributions on $\mathbb{N}$, parameterized by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)^T$. Here, for any $1 \leqslant l \leqslant r$, we assume that θ_l takes values from an open interval $\mathcal{I}_l$. For example, $\mathcal{I}_l$ may be $(0,1)$, $(0,\infty)$, or the whole real line $\mathbb{R}$. Write $\boldsymbol{\mathcal{I}} = \mathcal{I}_1 \times \cdots \times \mathcal{I}_r$ as the parameter space. Let p take value from an open interval $\mathcal{I}_0$. We use the following assumptions on regularity.

(A1). $f(k;\boldsymbol{\theta})$ is uniformly supported on positive integers. That is, for any $\boldsymbol{\theta} \in \boldsymbol{\mathcal{I}}$ and any integer $k \geqslant 1$, $f(k;\boldsymbol{\theta}) > 0$. Note that $f(0;\boldsymbol{\theta})$ may either be positive or zero, for different values of $\boldsymbol{\theta}$.

(A2). $f(k;\boldsymbol{\theta})$ is C^2 on $\boldsymbol{\theta}$. That is, for any fixed $k \in \mathbb{N}$, all the first order and second order partial derivatives of $f(k;\boldsymbol{\theta})$ are continuous on $\mathcal{I}$.

It is easy to verify that Pois(μ) and NB(μ,ν) both satisfy (A1) and (A2).

We now build the generalized linear model for the hurdle model f_p of f. For $0 \leqslant l \leqslant p$, let $g_l : \mathcal{I}_l \to \mathbb{R}$ be a link function such that g_l is invertible, g_l^{-1} is C^2, and $(g_l^{-1})'(t) > 0$ for all $t \in \mathbb{R}$. Most commonly used link functions satisfy these conditions. For example, the identity link $g_{\mathsf{id}}(t) = t$ on $\mathbb{R}$, the log link $g_{\mathsf{log}}(t) = \log(t)$ on $(0, \infty)$, the logit link $g_{\mathsf{logit}}(t) = \log \frac{t}{1-t}$ on $(0, 1)$, the probit link $g_{\mathsf{probit}}(t) = \Phi^{-1}(t)$ on $(0, 1)$ where $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-u^2/2} du$, and so on.

For the l'th parameter θ_l, we assume $\theta_l = g_l^{-1}(\boldsymbol{\beta}_l^T \boldsymbol{X}_l)$. Here, $\boldsymbol{\beta}_l = (\beta_l^1, \ldots, \beta_l^{d_l})^T \in \mathbb{R}^{d_l}$ is the coefficient vector, and $\boldsymbol{X}_l = (X_{l,1}, \ldots, X_{l,d_l})^T \in \mathbb{R}^{d_l}$ is the vector of covariates. For the model with intercept, one simply sets $X_{l,1} \equiv 1$. Note that for different l's, the covariate vectors $\boldsymbol{X}_l$ may share some common components. The hurdle parameter $p = g_0^{-1}(\boldsymbol{\beta}_0^T \boldsymbol{X}_0) \in \mathcal{I}_0$ is similarly defined on an open interval $\mathcal{I}_0 \subset (0, 1)$. Write $d = d_0 + d_1 + \cdots + d_r$. There are d parameters that specify the generalized linear hurdle model $f_p(k; \boldsymbol{\theta})$.

We model the grouped and right-censored counts by separating $\mathbb{N}$ into finite subsets, which we call groups. In particular, let N be the number of groups. We use a sequence of N integers $0 = l_1 < l_2 < \cdots < l_N < \infty$ to mark the boundaries of the groups. Write $l_{N+1} = \infty$. For $1 \leqslant k \leqslant N$, the k'th group is

$$G_k = \{i \in \mathbb{N} : l_k \leqslant i < l_{k+1}\}.$$

Denote $\mathcal{G} = \{l_k\}_{k=1}^{N+1}$ the grouping scheme. By grouping the probability masses of f_p, we obtain a categorical distribution on $\{1, \ldots, N\}$, of which the probability mass function $f_{\mathcal{G}}$ is defined by

$$f_{p,\mathcal{G}}(k; \boldsymbol{\theta}) = \sum_{l_k \leqslant i < l_{k+1}} f_p(i; \boldsymbol{\theta}), \qquad \text{for } 1 \leqslant k \leqslant N.$$

We now formulate the structure of sample with covariates. Write $D = \left\{(\boldsymbol{X}^i, Y_{\mathcal{G}}^i)\right\}_{i=1}^{n}$ as a sample of n independent observations drawn from the same distribution. Here $\boldsymbol{X}^i = (\boldsymbol{X}_0^i, \ldots, \boldsymbol{X}_r^i)^T$, $\boldsymbol{X}_0^i = (X_{0,1}^i, \ldots, X_{0,d_0}^i)^T$ is the covariate vector for p, and $\boldsymbol{X}_l^i = (X_{l,1}^i, \ldots, X_{l,d_l}^i)^T$ is the covariate vector for θ_l. In the literature, the covariate $\boldsymbol{X}^i$ can be modeled either as deterministic vectors (deterministic design), or as random from some unknown distribution (random design). We will discuss the design later. When $\boldsymbol{X}^i$ is given, the conditional distribution of $Y_{\mathcal{G}}^i \in \{1, \ldots, N\}$ is specified by the probability mass function $f_{p,\mathcal{G}}(k; \boldsymbol{\theta}^i)$, where $\boldsymbol{\theta}^i = (\theta_1^i, \ldots, \theta_r^i)^T$ with $\theta_l^i = g_l^{-1}(\boldsymbol{\beta}_l^T \boldsymbol{X}_l^i)$ and $p = g_0^{-1}(\boldsymbol{\beta}_0^T \boldsymbol{X}_0^i)$. So for estimating the coefficient vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_r)^T$, the log-likelihood function takes the form

$$\ell_{\mathcal{G}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f_{p,\mathcal{G}}(Y_{\mathcal{G}}^i, \boldsymbol{\theta}^i).$$

The grouping and right-censoring procedure usually spoils good algebraic properties of a distribution. For example, even if the original distribution f belongs to the exponential family, since the group number N is finite, the categorical distribution $f_{p,\mathcal{G}}$ of the corresponding hurdle model is in general not an exponential family distribution. Nonetheless, if f is smooth enough with respect to its parameters, and if smooth link functions are used, then the maximum likelihood estimator of the coefficient vector still enjoys asymptotic consistency and asymptotic normality. We characterize these properties in Theorem 3.1. Here we adopt the randomness assumption, that is, we assume that the predictors $\boldsymbol{X}^i$'s are drawn independently and identically from some unknown distribution. We point out that in the literature, the setting of fixed design parallels random design [25]. Here, the fixed design setting takes the predictors $\boldsymbol{X}^i$'s as deterministic variables.

Theorem 3.1. *Assume (A1), (A2), and that*

1. *The sample $D = \{(\boldsymbol{X}^i, Y_{\mathcal{G}}^i)\}_{i=1}^n$ is independently and identically drawn from a joint Borel probability distribution ρ on $\mathbb{R}^d \times \{1, \cdots, N\}$. Here, the marginal distribution ρ_X on $\mathbb{R}^d$ is supported on a compact set $\mathcal{X} \subset \mathbb{R}^d$, and for any $\boldsymbol{x} \in \mathcal{X}$, the conditional distribution $\rho(\cdot|\boldsymbol{x})$ on $\{1, \ldots, N\}$ is specified above through the grouped and right-censored hurdle model $f_{p,\mathcal{G}}$ with the coefficient vector $\boldsymbol{\beta}^*$, and the link functions $\{g_l\}_{l=0}^r$.*

2. *$f_{p,\mathcal{G}}(k;\boldsymbol{\theta})$ is C^2 with respect to $\boldsymbol{\beta}$, and g_l^{-1} is C^2 with $(g_l^{-1})' > 0$ everywhere for $0 \leqslant l \leqslant r$.*

3. *For any $0 \leqslant j \leqslant r$ and any $\boldsymbol{\xi} \in \mathbb{R}^{d_j}\backslash\{\mathbf{0}\}$,*
$$\mathbb{E}_{\boldsymbol{X}\sim\rho_j}\left[\langle \boldsymbol{X}, \boldsymbol{\xi}\rangle^2\right] > 0,$$
where ρ_j is the marginal distribution of ρ_X on $\mathbb{R}^{d_j}$ for j'th predictor vector of the model.

4. *The matrix $\mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta})$ is continuous and strictly positive definite at any $\boldsymbol{\theta} \in \boldsymbol{\mathcal{I}}$ and any $p \in \mathcal{I}_0$.*

Then, there exists a random integer n_1 and a sequence $\hat{\boldsymbol{\beta}}_n$ of random vectors, such that with the sample size $n \to \infty$, the following properties hold true.

(a). asymptotic existence, i.e., $\mathrm{Prob}\left(\nabla_{\boldsymbol{\beta}}\ell_{\mathcal{G}}(\hat{\boldsymbol{\beta}}_n) = 0 \text{ for all } n \geqslant n_1\right) = 1$*;*

(b). strong consistency, i.e., $\left\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\right\| \xrightarrow{a.s.} 0$*, as $n \to \infty$;*

(c). asymptotic normality, i.e.,
$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\right) \xrightarrow{in\ law} \mathcal{N}(\mathbf{0}, \mathbb{F}(\boldsymbol{\beta}^*)^{-1}),$$
where $\mathbb{F}(\boldsymbol{\beta}^*) = -\frac{1}{n}\mathbb{E}\left[\mathsf{Hessian}(\ell_{\mathcal{G}})(\boldsymbol{\beta}^*)\right]$ is the Fisher information matrix.

Theorem 3.1 is a direct corollary of Theorem A.1 in [35] and therefore we do not expand the proof.

3.2.3 The Computation of Fisher Information

We first consider the vanilla count model $f(k;\boldsymbol{\theta})$. Denote $f_{\mathcal{G}}$ the probability mass function on $\{1,\ldots,N\}$ obtained by grouping the probability mass of f according to the scheme $\mathcal{G}$. Define
$$f_{\mathcal{G}}(k;\boldsymbol{\theta}) = \sum_{i\in G_k} f(i;\boldsymbol{\theta}).$$

Denote $\mathbb{I}(f_{\mathcal{G}};\boldsymbol{\theta})$ the Fisher information matrix of size $r \times r$, of the distribution $f_{\mathcal{G}}$ at $\boldsymbol{\theta}$. Since $f_{\mathcal{G}}$ is a categorical distribution, an expectation with respect to $f_{\mathcal{G}}$ is just a finite sum, which is always interchangeable with partial differential operators, we have that for any $1 \leqslant i, j \leqslant N$,

$$\begin{aligned}\mathbb{I}(f_{\mathcal{G}},\boldsymbol{\theta})_{i,j} =& \mathbb{E}_{X\sim f_{\mathcal{G}}}\left[\left(\frac{\partial}{\partial\theta_i}\log f_{\mathcal{G}}(X;\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\theta_j}\log f_{\mathcal{G}}(X;\boldsymbol{\theta})\right)\right]\\ =& -\mathbb{E}_{X\sim f_{\mathcal{G}}}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f_{\mathcal{G}}(X;\boldsymbol{\theta})\right]\\ =& -\sum_{k=1}^{N} f_{\mathcal{G}}(k;\boldsymbol{\theta})\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f_{\mathcal{G}}(k;\boldsymbol{\theta})\\ =& \sum_{k=1}^{N}\frac{1}{f_{\mathcal{G}}(k;\boldsymbol{\theta})}\left(\frac{\partial}{\partial\theta_i}f_{\mathcal{G}}(k;\boldsymbol{\theta})\right)\left(\frac{\partial}{\partial\theta_j}f_{\mathcal{G}}(k;\boldsymbol{\theta})\right).\end{aligned}$$

We see that for computing $\mathbb{I}(f_{\mathcal{G}},\boldsymbol{\theta})$, we need only to compute $f_{\mathcal{G}}(k,\boldsymbol{\theta})$ and $\nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(k,\boldsymbol{\theta})$. In general, this can be achieved by computing

$$\sum_{i=a}^{b-1} f(i;\boldsymbol{\theta}), \qquad \text{and} \qquad \nabla_{\boldsymbol{\theta}}\sum_{i=a}^{b-1} f(i;\boldsymbol{\theta}),$$

for some integers (or infinity) $0 \leqslant a < b \leqslant \infty$.

For Poisson distributions $\text{Pois}(\mu)$, let $f^{\text{Pois}(\mu)}$ denote the probability mass function. Let $\lambda_i = e^{-\mu}\mu^i/i!$ for $i \geqslant 0$. For the sake of unified notation, let $\lambda_i = 0$ for $i < 0$ and let $\lambda_\infty = 0$. We have

$$\frac{d}{d\mu}\lambda_i = \lambda_{i-1} - \lambda_i, \qquad \text{for } -\infty < i \leqslant \infty.$$

So now, $\boldsymbol{\theta} = \mu \in \mathcal{I}_1 = (0,\infty)$ and

$$\frac{d}{d\mu}f_{\mathcal{G}}^{\text{Pois}(\mu)}(k;\mu) = \lambda_{l_k-1} - \lambda_{l_{k+1}-1}, \qquad \text{and}$$

$$\mathbb{I}(f_{\mathcal{G}}^{\text{Pois}(\mu)},\mu) = \sum_{k=1}^{N}\frac{\left(\lambda_{l_k-1} - \lambda_{l_{k+1}-1}\right)^2}{f_{\mathcal{G}}^{\text{Pois}(\mu)}(k;\mu)}.$$

For negative binomial distributions NB(μ, ν), let (recall that $\pi = \nu/(\mu+\nu)$)

$$\omega_i = \omega_i(\mu, \nu) = \frac{\Gamma(i+\nu)}{i!\Gamma(\nu)}\pi^{\nu}(1-\pi)^i.$$

For computing $\sum \omega_i$, we need the incomplete beta function

$$I_q(a,b) := \frac{1}{B(a,b)}\int_0^q t^{a-1}(1-t)^{b-1}dt,$$

where $0 \leqslant q \leqslant 1$, $a, b \in (0, \infty)$, and

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

is the beta function. One has $B(a,b) = B(b,a)$ and

$$\frac{\partial}{\partial a}B(a,b) = B(a,b)\left(\frac{\Gamma'(a)}{\Gamma(a)} - \frac{\Gamma'(a+b)}{\Gamma(a+b)}\right) = B(a,b)(\psi(a) - \psi(a+b)),$$

where $\psi(a) = \frac{d}{da}\log\Gamma(a)$ is the digamma function.

The mathematics of using incomplete beta function to represent and compute the probability mass function of NB(μ, ν) is well known. For example, this method has already been implemented in R (see, for example, [61] for a numerical algorithm for computing $I_q(a,b)$ with high precision). We include the derivation for the sake of completeness. For any integer $m \geqslant 0$, we have

$$\frac{\partial}{\partial \pi}I_\pi(\nu, m+1) = \frac{\Gamma(\nu+m+1)}{\Gamma(\nu)\Gamma(m+1)}\pi^{\nu-1}(1-\pi)^m.$$

Meanwhile,

$$\begin{aligned}
&\frac{\partial}{\partial \pi}\sum_{k=0}^{m}\frac{\Gamma(\nu+k)}{k!\Gamma(\nu)}\pi^{\nu}(1-\pi)^k \\
=&\sum_{k=0}^{m}\frac{\Gamma(\nu+k)}{k!\Gamma(\nu)}\left\{\nu\pi^{\nu-1}(1-\pi)^k + (1-\pi-1)k\pi^{\nu-1}(1-\pi)^{k-1}\right\} \\
=&\sum_{k=0}^{m}\frac{\Gamma(\nu+k+1)}{k!\Gamma(\nu)}\pi^{\nu-1}(1-\pi)^k - \sum_{k=1}^{m}\frac{\Gamma(\nu+k)}{(k-1)!\Gamma(\nu)}\pi^{\nu-1}(1-\pi)^{k-1} \\
=&\frac{\Gamma(\nu+m+1)}{\Gamma(\nu)\Gamma(m+1)}\pi^{\nu-1}(1-\pi)^m = \frac{\partial}{\partial \pi}I_\pi(\nu, m+1).
\end{aligned}$$

Since

$$\lim_{\pi \to 0^+} I_\pi(\nu, m+1) = \lim_{\pi \to 0^+} \sum_{k=0}^{m} \frac{\Gamma(\nu + k)}{k!\Gamma(\nu)} \pi^\nu (1-\pi)^k = 0,$$

one has

$$I_\pi(\nu, m+1) = \sum_{k=0}^{m} \omega_k.$$

For computing $\frac{\partial}{\partial \mu} \sum \omega_i$, consider

$$\begin{aligned}\frac{\partial}{\partial \mu}\omega_i =& \frac{\Gamma(i+\nu)}{i!\Gamma(\nu)}\pi^\nu(1-\pi)^i \left(\frac{\nu^2}{\pi}\frac{(-1)}{(\mu+\nu)^2} + \frac{i}{1-\pi}\frac{\mu+\nu-\mu}{(\mu+\nu)^2} \right) \\ =& \frac{\Gamma(i+\nu)}{i!\Gamma(\nu)}\pi^\nu(1-\pi)^i \left(-\pi + \frac{i\pi}{\mu} \right) \\ =& \frac{1}{\mu}\omega_i \pi (i - \mu) \\ =& \frac{1}{\mu}\left(i\omega_i - (i+1)\omega_{i+1} \right).\end{aligned}$$

Therefore we have that for any two integers $0 \leqslant a < b < \infty$,

$$\frac{\partial}{\partial \mu} \sum_{i=a}^{b-1} \omega_i = \frac{1}{\mu}\left(a\omega_a - b\omega_b \right). \tag{3.2}$$

One checks $\frac{\partial}{\partial \mu} \sum_{i=0}^{a-1} \omega_i = -\frac{1}{\mu} a\omega_a$ to find that the identity (3.2) also holds true for $b = \infty$ (here we use $\omega_\infty = 0$ for the sake of unified notation).

To our best knowledge, one has to take item-wise derivatives for computing $\frac{\partial}{\partial \nu} \sum \omega_i$ and there is no simpler method.

Now we start to discuss the representation of $\mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta})$ in terms of $\mathbb{I}(f_{\mathcal{G}}, \boldsymbol{\theta})$. The motivation is to develop a general numerical algorithm for computing $\mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta})$ with $\mathbb{I}(f_{\mathcal{G}}, \boldsymbol{\theta})$ as input. Note that $\mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta}) \in \mathbb{R}^{(r+1)\times(r+1)}$ and $\mathbb{I}(f_{\mathcal{G}}, \boldsymbol{\theta}) \in \mathbb{R}^{r \times r}$. In particular, we reserve the last row and the last column of $\mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta})$ for p. Denote the gradient $\nabla_{\boldsymbol{\theta}} h$ as a column vector for any C^1 function $h(\boldsymbol{\theta})$. Write $\mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta})[1:r, 1:r]$ the top left $r \times r$ sub-matrix of $\mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta})$, and write $\mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta})[1:r, r+1]$ the top right sub-matrix of size $r \times 1$. We summarize the representation in Theorem 3.2. We drop the vector $\boldsymbol{\theta}$ for light notations and write $f(k) = f(k; \boldsymbol{\theta})$, $f_{\mathcal{G}}(k) = f_{\mathcal{G}}(k; \boldsymbol{\theta})$, and $f_{p,\mathcal{G}}(k) = f_{p,\mathcal{G}}(k; \boldsymbol{\theta})$, respectively.

Theorem 3.2. *Write $R = (1 - f_{\mathcal{G}}(1))/(1 - f(0))$. We have*

$$\mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta})[1:r, 1:r] = \frac{p}{1-f(0)} \mathbb{I}(f_{\mathcal{G}}, \boldsymbol{\theta}) +$$

$$\frac{p}{(1-pR)(1-f(0))^2} \left[\nabla_{\boldsymbol{\theta}} f(0), \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1)\right] \begin{pmatrix} R & -1 \\ -1 & \frac{p-1+f(0)}{f_{\mathcal{G}}(1)} \end{pmatrix} \left[\nabla_{\boldsymbol{\theta}} f(0), \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1)\right]^T, \tag{3.3}$$

$$\mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta})_{r+1,r+1} = \frac{R}{p(1-pR)}, \quad \textit{and} \tag{3.4}$$

$$\mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta})[1:r, r+1] = \frac{1}{(1-pR)(1-f(0))} \left(R \nabla_{\boldsymbol{\theta}} f(0) - \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1)\right). \tag{3.5}$$

Proof. Recall that

$$\begin{aligned} f_{p,\mathcal{G}}(1) =& 1 - p + \frac{p}{1-f(0)} \sum_{j \in G_1, j \geqslant 2} f(j) \\ =& 1 - p + \frac{p(f_{\mathcal{G}}(1) - f(0))}{1 - f(0)} \\ =& 1 - Rp, \end{aligned}$$

and $R = 1$ when 0 is isolated (i.e., when 0 is separated out as a single group, $G_1 = \{0\}$). When 0 is not isolated, our assumption that f is supported on the whole $\mathbb{N}$ yields $0 < R < 1$. For $2 \leqslant k \leqslant N$,

$$f_{p,\mathcal{G}}(k) = \frac{p}{1-f(0)} f_{\mathcal{G}}(k).$$

For any $1 \leqslant i, j \leqslant r$, recall that $\frac{\partial}{\partial \theta_i} \log p = 0$,

$$\begin{aligned}
\mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta})_{i,j} =& -\sum_{k=1}^{N} f_{p,\mathcal{G}}(k) \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f_{p,\mathcal{G}}(k) \\
=& -(1-pR) \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log(1-pR) \\
& - \sum_{k=2}^{N} \frac{p}{1-f(0)} f_{\mathcal{G}}(k) \frac{\partial^2}{\partial \theta_i \partial \theta_j} \left(\log p + \log \frac{f_{\mathcal{G}}(k)}{1-f(0)} \right) \\
=:& J_1 + J_2 + J_3,
\end{aligned}$$

where J_1, J_2, and J_3 will be defined and calculated below. First,

$$\begin{aligned}
J_1 =& -(1-pR) \frac{\partial}{\partial \theta_i} \left(\frac{-p}{1-pR} \cdot \frac{\partial R}{\partial \theta_j} \right) \\
=& \frac{p^2}{1-pR} \frac{\partial R}{\partial \theta_i} \frac{\partial R}{\partial \theta_j} + p \frac{\partial^2 R}{\partial \theta_i \partial \theta_j}.
\end{aligned}$$

Next,

$$\begin{aligned}
J_2 =& -\frac{p}{1-f(0)} \sum_{k=2}^{N} f_{\mathcal{G}}(k) \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f_{\mathcal{G}}(k) \\
=& \frac{p}{1-f(0)} \mathbb{I}(f_{\mathcal{G}}, \boldsymbol{\theta})_{i,j} + \frac{p}{1-f(0)} f_{\mathcal{G}}(1) \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f_{\mathcal{G}}(1) \\
=& \frac{p}{1-f(0)} \mathbb{I}(f_{\mathcal{G}}, \boldsymbol{\theta})_{i,j} + \frac{p}{1-f(0)} \frac{\partial^2}{\partial \theta_i \partial \theta_j} f_{\mathcal{G}}(1) \\
& - \frac{p}{(1-f(0)) f_{\mathcal{G}}(1)} \frac{\partial f_{\mathcal{G}}(1)}{\partial \theta_i} \frac{\partial f_{\mathcal{G}}(1)}{\partial \theta_j}.
\end{aligned}$$

Then,

$$\begin{aligned}J_3 =&\frac{p}{1-f(0)}\sum_{k=2}^{N} f_{\mathcal{G}}(k)\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log(1-f(0))\\ =&p\frac{1-f_{\mathcal{G}}(1)}{1-f(0)}\frac{\partial}{\partial\theta_i}\left(\frac{-1}{1-f(0)}\cdot\frac{\partial f(0)}{\partial\theta_j}\right)\\ =&-\frac{pR}{1-f(0)}\frac{\partial^2 f(0)}{\partial\theta_i\partial\theta_j}-\frac{pR}{(1-f(0))^2}\frac{\partial f(0)}{\partial\theta_i}\frac{\partial f(0)}{\partial\theta_j}.\end{aligned}$$

We now revisit J_1. From

$$\begin{aligned}\frac{\partial R}{\partial\theta_i}=&-\frac{1}{1-f(0)}\frac{\partial f_{\mathcal{G}}(1)}{\partial\theta_i}+\frac{1-f_{\mathcal{G}}(1)}{(1-f(0))^2}\frac{\partial f(0)}{\partial\theta_i}\\ =&\frac{-1}{1-f(0)}\frac{\partial f_{\mathcal{G}}(1)}{\partial\theta_i}+\frac{R}{1-f(0)}\frac{\partial f(0)}{\partial\theta_i},\end{aligned}\tag{3.6}$$

we have

$$\begin{aligned}\frac{\partial^2 R}{\partial\theta_i\partial\theta_j}=&\frac{-1}{(1-f(0))^2}\frac{\partial f(0)}{\partial\theta_j}\frac{\partial f_{\mathcal{G}}(1)}{\partial\theta_i}-\frac{1}{1-f(0)}\frac{\partial^2 f_{\mathcal{G}}(1)}{\partial\theta_i\partial\theta_j}\\ &+\frac{R}{(1-f(0))^2}\frac{\partial f(0)}{\partial\theta_i}\frac{\partial f(0)}{\partial\theta_j}+\frac{R}{1-f(0)}\frac{\partial^2 f(0)}{\partial\theta_i\partial\theta_j}\\ &+\frac{1}{1-f(0)}\frac{\partial f(0)}{\partial\theta_i}\left(\frac{R}{1-f(0)}\frac{\partial f(0)}{\partial\theta_j}-\frac{1}{1-f(0)}\frac{\partial f_{\mathcal{G}}(1)}{\partial\theta_j}\right)\\ =&\frac{-1}{1-f(0)}\frac{\partial^2 f_{\mathcal{G}}(1)}{\partial\theta_i\partial\theta_j}-\frac{1}{(1-f(0))^2}\left(\frac{\partial f(0)}{\partial\theta_i}\frac{\partial f_{\mathcal{G}}(1)}{\partial\theta_j}+\frac{\partial f_{\mathcal{G}}(1)}{\partial\theta_i}\frac{\partial f(0)}{\partial\theta_j}\right)\\ &+\frac{R}{1-f(0)}\frac{\partial^2 f(0)}{\partial\theta_i\partial\theta_j}+\frac{2R}{(1-f(0))^2}\frac{\partial f(0)}{\partial\theta_i}\frac{\partial f(0)}{\partial\theta_j}.\end{aligned}$$

Therefore,

$$\begin{aligned}J_1 =&\frac{p^2}{(1-pR)(1-f(0))^2}\frac{\partial f_{\mathcal{G}}(1)}{\partial\theta_i}\frac{\partial f_{\mathcal{G}}(1)}{\partial\theta_j}+\frac{p^2R^2}{(1-pR)(1-f(0))^2}\frac{\partial f(0)}{\partial\theta_i}\frac{\partial f(0)}{\partial\theta_j}\\ &-\left(\frac{p^2R}{(1-pR)(1-f(0))^2}+\frac{p}{(1-f(0))^2}\right)\left(\frac{\partial f(0)}{\partial\theta_i}\frac{\partial f_{\mathcal{G}}(1)}{\partial\theta_j}+\frac{\partial f_{\mathcal{G}}(1)}{\partial\theta_i}\frac{\partial f(0)}{\partial\theta_j}\right)\\ &-\frac{p}{1-f(0)}\frac{\partial^2 f_{\mathcal{G}}(1)}{\partial\theta_i\partial\theta_j}+\frac{pR}{1-f(0)}\frac{\partial^2 f(0)}{\partial\theta_i\partial\theta_j}+\frac{2pR}{(1-f(0))^2}\frac{\partial f(0)}{\partial\theta_i}\frac{\partial f(0)}{\partial\theta_j}.\end{aligned}$$

We combine the above calculation to obtain

$$
\begin{aligned}
\mathbb{I}(f_{p,\mathcal{G}},\boldsymbol{\theta})_{i,j} =& \frac{p}{1-f(0)}\mathbb{I}(f_{\mathcal{G}},\boldsymbol{\theta})_{i,j} + \frac{p(p-1+f(0))}{f_{\mathcal{G}}(1)(1-pR)(1-f(0))^2}\frac{\partial f_{\mathcal{G}}(1)}{\partial\theta_i}\frac{\partial f_{\mathcal{G}}(1)}{\partial\theta_j} \\
&+ \frac{pR}{(1-pR)(1-f(0))^2}\frac{\partial f(0)}{\partial\theta_i}\frac{\partial f(0)}{\partial\theta_j} \\
&- \frac{p}{(1-pR)(1-f(0))^2}\left(\frac{\partial f(0)}{\partial\theta_i}\frac{\partial f_{\mathcal{G}}(1)}{\partial\theta_j} + \frac{\partial f_{\mathcal{G}}(1)}{\partial\theta_i}\frac{\partial f(0)}{\partial\theta_j}\right),
\end{aligned}
$$

which proves (3.3).

For $\mathbb{I}(f_{p,\mathcal{G}},\boldsymbol{\theta})_{r+1,r+1}$, we have

$$
\begin{aligned}
\mathbb{I}(f_{p,\mathcal{G}},\boldsymbol{\theta})_{r+1,r+1} =& -\sum_{k=1}^{N} f_{p,\mathcal{G}}(k)\frac{\partial^2}{\partial p^2}\log f_{p,\mathcal{G}}(k) \\
=& -(1-pR)\frac{\partial^2}{\partial p^2}\log(1-pR) \\
&- \sum_{k=2}^{N}\frac{p}{1-f(0)}f_{\mathcal{G}}(k)\frac{\partial^2}{\partial p^2}\left(\log p + \log\frac{f_{\mathcal{G}}(k)}{1-f(0)}\right) \\
=& \frac{R^2}{1-pR} + \frac{1-f_{\mathcal{G}}(1)}{p(1-f(0))} \\
=& R\left(\frac{R}{1-pR}+\frac{1}{p}\right) \\
=& \frac{R}{p(1-pR)}.
\end{aligned}
$$

For $\mathbb{I}(f_{p,\mathcal{G}},\boldsymbol{\theta})[1:r,r+1]$, recall (3.6). Now the cross terms in the sum $\sum_{k=2}^{N}$ are all zero. Let $1 \leqslant i \leqslant r$ to obtain

$$\begin{aligned}
\mathbb{I}(f_{p,\mathcal{G}},\boldsymbol{\theta})_{i,r+1} =& - \sum_{k=1}^{N} f_{p,\mathcal{G}}(k) \frac{\partial^2}{\partial p \partial \theta_i} \log f_{p,\mathcal{G}}(k) \\
=& - (1-pR) \frac{\partial^2}{\partial p \partial \theta_i} \log(1-pR) \\
& - \sum_{k=2}^{N} \frac{p}{1-f(0)} f_{\mathcal{G}}(k) \frac{\partial^2}{\partial p \partial \theta_i} \left(\log p + \log \frac{f_{\mathcal{G}}(k)}{1-f(0)} \right) \\
=& - (1-pR) \frac{\partial}{\partial \theta_i} \frac{-R}{1-pR} \\
=& (1-pR) \left(\frac{1-pR+pR}{(1-pR)^2} \right) \frac{\partial R}{\partial \theta_i} \\
=& \frac{1}{(1-pR)(1-f(0))} \left(R \frac{\partial f(0)}{\partial \theta_i} - \frac{\partial f_{\mathcal{G}}(1)}{\partial \theta_i} \right).
\end{aligned}$$

The proof is complete. □

The following corollary is obtained by noting that when $G_1 = \{0\}$, we have $f(0) = f_{\mathcal{G}}(1)$ and $R = 1$. When 0 is isolated, $f_{p,\mathcal{G}}(1) = 1 - p$. The representation of $\mathbb{I}(f_{p,\mathcal{G}},\boldsymbol{\theta})$ by $\mathbb{I}(f_{\mathcal{G}},\boldsymbol{\theta})$ has a simpler form.

Corollary 3.1. *When* 0 *is isolated, that is, when* $G_1 = \{0\}$ *for the grouping scheme* $\mathcal{G}$, *we have*

$$\mathbb{I}(f_{p,\mathcal{G}},\boldsymbol{\theta}) = \begin{bmatrix} \dfrac{p}{1-f(0)} \left(\mathbb{I}(f_{\mathcal{G}},\boldsymbol{\theta}) - \dfrac{\nabla_{\boldsymbol{\theta}} f(0) \nabla_{\boldsymbol{\theta}} f(0)^T}{f(0)(1-f(0))} \right) & 0 \\ 0 & \dfrac{1}{p(1-p)} \end{bmatrix} \tag{3.7}$$

Proof. In (3.3), we substitute $f_{\mathcal{G}}(1) = f(0)$, $\nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1) = \nabla_{\boldsymbol{\theta}} f(0)$, and $R = 1$ to obtain

$$\begin{aligned}
&\frac{p}{(1-pR)(1-f(0))^2} \left[\nabla_{\boldsymbol{\theta}} f(0), \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1)\right] \begin{pmatrix} R & -1 \\ -1 & \frac{p-1+f(0)}{f_{\mathcal{G}}(1)} \end{pmatrix} \left[\nabla_{\boldsymbol{\theta}} f(0), \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1)\right]^T \\
=&\frac{p}{(1-p)(1-f(0))^2} \left(-1 + \frac{p-1+f(0)}{f(0)}\right) \nabla_{\boldsymbol{\theta}} f(0) \nabla_{\boldsymbol{\theta}} f(0)^T \\
=&-\frac{p}{f(0)(1-f(0))^2} \nabla_{\boldsymbol{\theta}} f(0) \nabla_{\boldsymbol{\theta}} f(0)^T.
\end{aligned}$$

This proves the top left corner of the matrix in (3.7). The rest part of the matrix is evident. The proof is complete. □

As suggested by Theorem 3.1, it is important to make sure that $\mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta})$ is strictly positive definite, during the design of the grouping scheme $\mathcal{G}$. To develop the theory, we write $\mathcal{G}_{\dagger}^0$ the grouping scheme that has zero isolated and the rest integers in $\mathbb{N}$ put as the other group. Namely,

$$\mathcal{G}_{\dagger}^0 = \{0, 1, \infty\}.$$

Therefore,

$$\begin{aligned}
f_{\mathcal{G}_{\dagger}^0}(1) =& f(0), \\
f_{\mathcal{G}_{\dagger}^0}(2) =& 1 - f(0).
\end{aligned}$$

The Fisher information matrix is computed by

$$\begin{aligned}
\mathbb{I}(f_{\mathcal{G}_{\dagger}^0}, \boldsymbol{\theta}) =& \sum_{k=1}^{2} \frac{1}{f_{\mathcal{G}_{\dagger}^0}(k)} \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}_{\dagger}^0}(k) \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}_{\dagger}^0}(k)^T \\
=& \left(\frac{1}{f(0)} + \frac{1}{1-f(0)}\right) \nabla_{\boldsymbol{\theta}} f(0) \nabla_{\boldsymbol{\theta}} f(0)^T \\
=& \frac{1}{f(0)(1-f(0))} \nabla_{\boldsymbol{\theta}} f(0) \nabla_{\boldsymbol{\theta}} f(0)^T.
\end{aligned}$$

This observation links $\mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta})$ and $\mathbb{I}(f_{\mathcal{G}}, \boldsymbol{\theta})$ by

$$\mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta}) = \begin{bmatrix} \frac{p}{1-f(0)} \left(\mathbb{I}(f_{\mathcal{G}}, \boldsymbol{\theta}) - \mathbb{I}(f_{\mathcal{G}_{\dagger}^0}, \boldsymbol{\theta})\right) & 0 \\ 0 & \frac{1}{p(1-p)} \end{bmatrix}.$$

More importantly, we have the following characterization of $\mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta})$ when 0 is isolated in $\mathcal{G}$. The theorem follows directly from Corollary 3.1.

Theorem 3.3. *When* 0 *is isolated, that is, when* $G_1 = \{0\}$ *for the grouping scheme* $\mathcal{G}$, $\mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta})$ *is strictly positive definite if and only if* $\mathbb{I}(f_{\mathcal{G}}, \boldsymbol{\theta}) - \mathbb{I}(f_{\mathcal{G}^0_\dagger}, \boldsymbol{\theta})$ *is strictly positive definite.* □

Before we move on, we would prepare some notations.

- For any grouping scheme $\mathcal{G}$, denote $|\mathcal{G}|$ the number of groups contained in $\mathcal{G}$;
- For any grouping scheme $\mathcal{G}$ with $|\mathcal{G}| \geqslant 2$, let $\mathcal{G}_\dagger$ denote the grouping scheme obtained by merging all but the first group of $\mathcal{G}$ as one. That is, if $N = |\mathcal{G}| \geqslant 2$ with $\mathcal{G} = \{0 = l_1, l_2, \ldots, l_{N+1} = \infty\}$, then $|\mathcal{G}_\dagger| = 2$ with $\mathcal{G}_\dagger = \{l_1, l_2, l_{N+1}\}$. So, if $N = 2$, $\mathcal{G} = \mathcal{G}_\dagger$.

The following theorem provides a characterization for the structure of general grouped and right censored hurdle models. It covers Theorem 3.3 as a direct consequence.

Theorem 3.4. *Let* $\mathcal{G}$ *be a grouping scheme. Here, the integer zero may either be isolated or not, so the first group* G_1 *of* $\mathcal{G}$ *may contain either only zero, or more integers. Then,* $\mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta})$ *is strictly positive definite if and only if* $\mathbb{I}(f_{\mathcal{G}}, \boldsymbol{\theta}) - \mathbb{I}(f_{\mathcal{G}_\dagger}, \boldsymbol{\theta})$ *is strictly positive definite.*

The proof is organized as follows. We shall write

$$P\mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta})P^T = \left[\begin{array}{cc} \dfrac{p}{1-f(0)} \left(\mathbb{I}(f_{\mathcal{G}}, \boldsymbol{\theta}) - \mathbb{I}(f_{\mathcal{G}_\dagger}, \boldsymbol{\theta})\right) & \mathbf{0} \\ \mathbf{0}^T & \dfrac{R}{p(1-pR)} \end{array} \right], \tag{3.8}$$

where P is an invertible matrix. Then the proof is completed by noting the facts that $\frac{p}{1-f(0)} > 0$ and $\frac{R}{p(1-pR)} > 0$.

Proof of Theorem 3.4. We use

$$P = \left[\begin{array}{cc} I & -v_P \\ \mathbf{0}^T & 1 \end{array} \right],$$

where $I \in \mathbb{R}^{r\times r}$ is the identity matrix, $\mathbf{0} \in \mathbb{R}^{r\times 1}$ is a zero vector, and

$$v_P = \frac{p(1-pR)}{R} \frac{1}{(1-pR)(1-f(0))} \left(R\nabla_{\boldsymbol{\theta}} f(0) - \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1) \right).$$

Obviously P is invertible. Note that for $A \in \mathbb{R}^{r\times r}$, $b \in \mathbb{R}^{r\times 1}$ and $c \in \mathbb{R}$, we have

$$\begin{aligned}
&\left[\begin{array}{cc} I & -v_P \\ \mathbf{0}^T & 1 \end{array} \right] \left[\begin{array}{cc} A & b \\ b^T & c \end{array} \right] \left[\begin{array}{cc} I & \mathbf{0} \\ -v_P^T & 1 \end{array} \right] \\
=&\left[\begin{array}{cc} A - v_P b^T & b - cv_P \\ b^T & c \end{array} \right] \left[\begin{array}{cc} I & \mathbf{0} \\ -v_P^T & 1 \end{array} \right] \\
=&\left[\begin{array}{cc} A - v_P b^T - bv_P^T + cv_P v_P^T & b - cv_P \\ b^T - cv_P^T & c \end{array} \right].
\end{aligned}$$

We substitute the matrix

$$\left[\begin{array}{cc} A & b \\ b^T & c \end{array} \right]$$

by $\mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta})$ to obtain that first,

$$\begin{aligned}
&\left(P\mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta})P^T\right)[1:r, 1:r] \\
=&\frac{p}{1-f(0)}\mathbb{I}(f_{\mathcal{G}}, \boldsymbol{\theta}) + \frac{pR}{(1-pR)(1-f(0))^2}\nabla_{\boldsymbol{\theta}} f(0)\nabla_{\boldsymbol{\theta}} f(0)^T \\
&- \frac{p}{(1-pR)(1-f(0))^2}\left(\nabla_{\boldsymbol{\theta}} f(0)\nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1)^T + \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1)\nabla_{\boldsymbol{\theta}} f(0)^T\right) \\
&+ \frac{p(p-1+f(0)}{(1-pR)(1-f(0))^2 f_{\mathcal{G}}(1)}\nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1)\nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1)^T \\
&- \frac{p(1-pR)}{R(1-pR)^2(1-f(0))^2}\left(R\nabla_{\boldsymbol{\theta}} f(0) - \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1)\right)\left(R\nabla_{\boldsymbol{\theta}} f(0) - \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1)\right)^T.
\end{aligned}$$

In the above equation, the coefficient of $\nabla_{\boldsymbol{\theta}} f(0)\nabla_{\boldsymbol{\theta}} f(0)^T$ is

$$\frac{pR}{(1-pR)(1-f(0))^2} - \frac{pR^2(1-pR)}{R(1-pR)^2(1-f(0))^2} = 0.$$

The coefficient of $\nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1) \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1)^T$ (excluding the component in $\frac{p}{1-f(0)} \mathbb{I}(f_{\mathcal{G}}, \boldsymbol{\theta})$) is (recall that $R = (1 - f_{\mathcal{G}}(1))/(1 - f(0))$)

$$\frac{p(p-1+f(0))}{(1-pR)(1-f(0))^2 f_{\mathcal{G}}(1)} - \frac{p}{R(1-pR)(1-f(0))^2} \tag{3.9}$$

$$= \frac{p^2 R - p(1-f_{\mathcal{G}}(1)) - p f_{\mathcal{G}}(1)}{R(1-pR)(1-f(0))^2 f_{\mathcal{G}}(1)} \tag{3.10}$$

$$= -\frac{p}{(1-f(0)) f_{\mathcal{G}}(1)(1-f_{\mathcal{G}}(1))}. \tag{3.11}$$

The coefficients of $\nabla_{\boldsymbol{\theta}} f(0) \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1)^T$ and $\nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1) \nabla_{\boldsymbol{\theta}} f(0)^T$ are the same,

$$\frac{-p}{(1-pR)(1-f(0))^2} + \frac{pR}{R(1-pR)(1-f(0))^2} = 0.$$

Therefore,

$$\begin{aligned} &\left(P \mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta}) P^T\right)[1:r, 1:r] \\ &= \frac{p}{1-f(0)} \mathbb{I}(f_{\mathcal{G}}, \boldsymbol{\theta}) - \frac{p}{(1-f(0)) f_{\mathcal{G}}(1)(1-f_{\mathcal{G}}(1))} \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1) \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1)^T. \end{aligned}$$

Recall that since $\mathcal{G}$ and $\mathcal{G}_\dagger$ share the same first group, $f_{\mathcal{G}}(1) = f_{\mathcal{G}_\dagger}(1)$. Since $\mathcal{G}_\dagger$ has only two groups, $f_{\mathcal{G}_\dagger}(2) = 1 - f_{\mathcal{G}_\dagger}(1) = 1 - f_{\mathcal{G}}(1)$. We have

$$\begin{aligned} \mathbb{I}(f_{\mathcal{G}_\dagger}, \boldsymbol{\theta}) &= \sum_{k=1}^{2} \frac{1}{f_{\mathcal{G}_\dagger}(k)} \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}_\dagger}(k) \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}_\dagger}(k)^T \\ &= \left(\frac{1}{f_{\mathcal{G}}(1)} + \frac{1}{1-f_{\mathcal{G}}(1)}\right) \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1) \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1)^T \\ &= \frac{1}{f_{\mathcal{G}}(1)(1-f_{\mathcal{G}}(1))} \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1) \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1)^T. \end{aligned}$$

This yields

$$\left(P \mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta}) P^T\right)[1:r, 1:r] = \frac{p}{1-f(0)} \left(\mathbb{I}(f_{\mathcal{G}}, \boldsymbol{\theta}) - \mathbb{I}(f_{\mathcal{G}_\dagger}, \boldsymbol{\theta})\right).$$

Next,

$$\left(P \mathbb{I}(f_{p,\mathcal{G}}, \boldsymbol{\theta}) P^T\right)[1:r, r+1]$$

$$= \frac{1}{(1-pR)(1-f(0))} \left(R \nabla_{\boldsymbol{\theta}} f(0) - \nabla_{\boldsymbol{\theta}} f_{\mathcal{G}}(1)\right) - \frac{R}{p(1-pR)} v_P = \mathbf{0}.$$

Similarly,

$$\left(P\mathbb{I}(f_{p,\mathcal{G}},\boldsymbol{\theta})P^T\right)[r+1,1:r]=\mathbf{0}^T.$$

So, (3.8) is proved, and the proof is complete. □

Let $\mathcal{G}$ and $\mathcal{G}'$ be two grouping schemes. We say that $\mathcal{G}'$ is *finer* than $\mathcal{G}$ and write $\mathcal{G}' \succ \mathcal{G}$ or $\mathcal{G} \prec \mathcal{G}'$, if $\mathcal{G}'$ is obtained by dividing one or several groups of $\mathcal{G}$ to smaller groups respectively. By this operation, each group in $\mathcal{G}'$ is contained entirely in one group in $\mathcal{G}$, $\mathcal{G} \subset \mathcal{G}'$ and $|\mathcal{G}'| \geqslant |\mathcal{G}| + 1$.

The following theorem is from [30].

Theorem 3.5 ([30])**.** *Consider two grouping schemes $\mathcal{G}$ and $\mathcal{G}'$. Let $0 < \mu < \infty$. Then $\mathcal{G} \prec \mathcal{G}'$ implies $\mathbb{I}(f_{\mathcal{G}}^{\mathrm{Pois}(\mu)}, \mu) < \mathbb{I}(f_{\mathcal{G}'}^{\mathrm{Pois}(\mu)}, \mu)$.*

As a direct consequence, we have the following corollary.

Corollary 3.2. *Let $|\mathcal{G}| \geqslant 3$. Then $\mathbb{I}(f_{p,\mathcal{G}}^{\mathrm{Pois}(\mu)}, \mu)$ is strictly positive definite for any $0 < p < 1$ and $0 < \mu < \infty$.*

Corollary 3.2 shows that with the other assumptions in Theorem 3.1, one needs only 3 groups for an asymptotically consistent parameter inference for generalized linear models with hurdle Poisson distributions.

For negative binomial distributions, we have not found similar results. We would leave the topic as future research.

3.3 Discussion and Conclusions

In this work, we explored some inspiring and interesting properties of grouped and right-censored hurdle models. In particular, we showed that under mild conditions the maximum likelihood estimator of grouped and right-censored hurdle models is asymptotically consistent and normal. We discussed the computational issues of Fisher information, and established the relations between the Fisher information matrices of grouped and right-censored models and the corresponding hurdle models, with the motivation of developing a stand-alone algorithm for grouped and right-censored hurdle model inference that is independent of specific count distribution families. As a consequence, we developed a simple sufficient and necessary condition for the Fisher information matrix of grouped and right-censored hurdle model to be strictly positive definite. Therefore, we now see that one needs only three groups for Poisson distributions to achieve such strictly positive definiteness.

Bibliography

[1] Alan Agresti. *Categorical data analysis*, volume 482. John Wiley & Sons, 2003.

[2] Akiko Aizawa. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65, 2003.

[3] Nikolaos Aletras and Mark Stevenson. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*, pages 13–22, 2013.

[4] Paul D Allison and Richard P Waterman. Fixed–effects negative binomial regression models. *Sociological methodology*, 32(1):247–265, 2002.

[5] Sylvain Arlot, Alain Celisse, et al. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

[6] David N Barron. The analysis of count data: Overdispersion and autocorrelation. *Sociological methodology*, pages 179–220, 1992.

[7] Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *J. Complexity*, 23(1):52–72, 2007.

[8] Andrea L. Bertozzi, Xiyang Luo, Andrew M. Stuart, and Konstantinos C. Zygalakis. Uncertainty quantification in graph-based classification of high dimensional data. *SIAM/ASA J. Uncertain. Quantif.*, 6(2):568–595, 2018.

[9] Rajendra Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.

[10] Rajendra Bhatia and Ludwig Elsner. The Hoffman-Wielandt inequality in infinite dimensions. *Proc. Indian Acad. Sci. Math. Sci.*, 104(3):483–494, 1994.

[11] Gilles Blanchard and Nicole Krämer. Optimal learning rates for kernel conjugate gradient regression. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 226–234. Curran Associates, Inc., 2010.

[12] Gilles Blanchard and Nicole Krämer. Convergence rates of kernel conjugate gradient for random design regression. *Anal. Appl.*, 14(6):763–794, 2016.

[13] David M Blei. Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1:203–232, 2014.

[14] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[15] Kurt Brännäs. Limited dependent poisson regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 41(4):413–423, 1992.

[16] A Colin Cameron and Pravin K Trivedi. *Regression analysis of count data*, volume 53. Cambridge university press, 2013.

[17] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7(3):331–368, 2007.

[18] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, pages 288–296, 2009.

[19] Xiangyu Chang, Shao-Bo Lin, and Ding-Xuan Zhou. Distributed semi-supervised learning with kernel ridge regression. *J. Mach. Learn. Res.*, 18:Paper No. 46, 22, 2017.

[20] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *J. Mach. Learn. Res.*, 12:1069–1109, 2011.

[21] Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21(1):5–30, 2006.

[22] Felipe Cucker and Ding-Xuan Zhou. *Learning theory: an approximation theory viewpoint*, volume 24 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2007. With a foreword by Stephen Smale.

[23] Russell Davidson and James G MacKinnon. *Econometric theory and methods*, volume 5. Oxford University Press New York, 2004.

[24] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[25] Ludwig Fahrmeir and Heinz Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.*, 13(1):342–368, 1985.

[26] R. A. Fisher. The negative binomial distribution. *Annals of Eugenics*, 11:182–187, 1941.

[27] John Fox. *Applied regression analysis and generalized linear models.* Sage Publications, 2015.

[28] Qiang Fu, Xin Guo, and Kenneth C. Land. A Poisson-multinomial mixture approach to grouped and right-censored counts. *Comm. Statist. Theory Methods*, 47(2):427–447, 2018.

[29] Qiang Fu, Xin Guo, and Kenneth C. Land. Optimizing count responses in surveys: A machine-learning approach. *Sociological Methods & Research*, 49(3):637–671, 2020.

[30] Qiang Fu, Xin Guo, and Kenneth C. Land. Optimizing count responses in surveys: A machine-learning approach. *Sociological Methods & Research*, 49(3):637–671, 2020.

[31] Qiang Fu, Kenneth C Land, and Vicki L Lamb. Bullying victimization, socioeconomic status and behavioral characteristics of 12th graders in the united states, 1989 to 2009: repetitive trends and persistent risk differentials. *Child Indicators Research*, 6(1):1–21, 2013.

[32] Qiang Fu, Kenneth C Land, and Vicki L Lamb. Violent physical bullying victimization at school: has there been a recent increase in exposure or intensity? an age-period-cohort analysis in the united states, 1991 to 2012. *Child Indicators Research*, 9(2):485–513, 2016.

[33] Qiang Fu and Qiang Ren. Educational inequality under china's rural–urban divide: The hukou system and return to education. *Environment and Planning A*, 42(3):592–610, 2010.

[34] Qiang Fu, Cary Wu, Heqing Liu, Zhilei Shi, and Jiaxin Gu. Live like mosquitoes: Hukou, rural–urban disparity, and depression. *Chinese Journal of Sociology*, 4(1):56–78, 2018.

[35] Qiang Fu, Tian-yi Zhou, and Xin Guo. Modified poisson regression analysis of grouped and right-censored counts. Manuscript submitted for publication, 2020.

[36] Wenjiang J Fu. Ridge estimator in singulah oesiun with application to age-period-cohort analysis of disease rates. *Communications in statistics-Theory and Methods*, 29(2):263–278, 2000.

[37] Wenjiang J Fu. A smoothing cohort model in age–period–cohort analysis with applications to homicide arrest rates and lung cancer mortality rates. *Sociological methods & research*, 36(3):327–361, 2008.

[38] Wenjiang J. Fu and Peter Hall. Asymptotic properties of estimators in age-period-cohort analysis. *Statist. Probab. Lett.*, 76(17):1925–1929, 2006.

[39] Jing Gao and Jun Zhang. Clustered SVD strategies in latent semantic indexing. *Information Processing & Management*, 41(5):1051–1063, 2005.

[40] Eric Gaussier and Cyril Goutte. Relation between plsa and nmf and implications. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 601–602. ACM, 2005.

[41] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[42] Zoubin Ghahramani, Geoffrey E Hinton, et al. The EM algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto, 1996.

[43] William H Greene. *Econometric analysis*. Upper Saddle River, N.J.; Prentice Hall, 5th ed.,international ed edition, 2003.

[44] Shulamith T. Gross and Tze Leung Lai. Nonparametric estimation and regression analysis with left-truncated and right-censored data. *J. Amer. Statist. Assoc.*, 91(435):1166–1180, 1996.

[45] Xin Guo, Ting Hu, and Qiang Wu. Distributed minimum error entropy algorithms. Preprint, 2019.

[46] Xin Guo and Ding-Xuan Zhou. An empirical feature-based learning algorithm producing sparse approximations. *Appl. Comput. Harmon. Anal.*, 32(3):389–400, 2012.

[47] Zheng-Chu Guo, Shao-Bo Lin, and Ding-Xuan Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 33(7):074009, 29, 2017.

[48] Zheng-Chu Guo, Lei Shi, and Qiang Wu. Learning theory of distributed regression with bias corrected regularization kernel network. *J. Mach. Learn. Res.*, 18:Paper No. 118, 25, 2017.

[49] Daniel B. Hall. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, 56(4):1030–1039, 2000.

[50] Fumio Hayashi. *Econometrics*. Princeton University Press, 2000.

[51] A. J. Hoffman and H. W. Wielandt. The variation of the spectrum of a normal matrix. *Duke Math. J.*, 20:37–39, 1953.

[52] Matthew Hoffman, Francis R Bach, and David M Blei. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 856–864, 2010.

[53] Ting Hu, Jun Fan, Qiang Wu, and Ding-Xuan Zhou. Regularization schemes for minimum error entropy principle. *Anal. Appl. (Singap.)*, 13(4):437–455, 2015.

[54] Ian Jolliffe. *Principal Component Analysis*. Springer, 2011.

[55] Tosio Kato. Variation of discrete spectra. *Comm. Math. Phys.*, 111(3):501–504, 1987.

[56] Diane Lambert. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

[57] Kenneth C Land, Patricia L McCall, and Daniel S Nagin. A comparison of poisson, negative binomial, and semiparametric mixed poisson regression models: With empirical applications to criminal careers data. *Sociological Methods & Research*, 24(4):387–442, 1996.

[58] Kenneth C Land, Emma Zang, Qiang Fu, Xin Guo, Sun Y Jeon, and Eric N Reither. Playing with the rules and making misleading statements: a response to luo, hodges, winship, and powers. *American Journal of Sociology*, 122(3):962–973, 2016.

[59] Thomas K Landauer, Peter W Foltz, and Darrell Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284, 1998.

[60] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788, 1999.

[61] Russell V. Lenth. Algorithm as 226: Computing noncentral beta probabilities. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36(2):241–244, 1987.

[62] Si-ming Li, Yushu Zhu, and Limei Li. Neighborhood type, gatedness, and residential experiences in chinese cities: A study of guangzhou. *Urban geography*, 33(2):237–255, 2012.

[63] Shao-Bo Lin, Xin Guo, and Ding-Xuan Zhou. Distributed learning with regularized least squares. *J. Mach. Learn. Res.*, 18:Paper No. 92, 31, 2017.

[64] Shao-Bo Lin and Ding-Xuan Zhou. Distributed kernel-based gradient descent algorithms. *Constr. Approx.*, 47(2):249–276, 2018.

[65] G. Little and J. B. Reade. Eigenvalues of analytic kernels. *SIAM J. Math. Anal.*, 15(1):133–136, 1984.

[66] Jin Liu, Can Yang, Yuling Jiao, and Jian Huang. ssLasso: a summary-statistic-based regression using Lasso. Preprint, 2017.

[67] J Scott Long and Jeremy Freese. *Regression models for categorical dependent variables using Stata*. Stata press, 2006.

[68] Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *Association for Computational Linguistics*, 2004.

[69] Lawrence C Marsh and David R Cormier. *Spline regression models*. Number 137 in 07. Sage, 2001.

[70] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics, 2011.

[71] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*, volume 821. John Wiley & Sons, 2012.

[72] John Neter, Michael H Kutner, Christopher J Nachtsheim, and William Wasserman. *Applied linear statistical models*. Irwin Chicago, 1996.

[73] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics, 2010.

[74] Normand Péladeau and Elnaz Davoodi. Comparison of latent dirichlet modeling and factor analysis for topic extraction: A lesson of history. In *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[75] Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *Ann. Probab.*, 22(4):1679–1706, 1994.

[76] Rafal Raciborski. Right-censored poisson regression model. *The Stata Journal*, 11(1):95–105, 2011.

[77] JNK Rao and AJ Scott. A simple method for analysing overdispersion in clustered poisson data. *Statistics in medicine*, 18(11):1373–1385, 1999.

[78] J. B. Reade. Eigenvalues of positive definite kernels. II. *SIAM J. Math. Anal.*, 15(1):137–142, 1984.

[79] Eric N Reither, Kenneth C Land, Sun Y Jeon, Daniel A Powers, Ryan K Masters, Hui Zheng, Melissa A Hardy, Katherine M Keyes, Qiang Fu, Heidi A Hanson, et al. Clarifying hierarchical age–period–cohort models: A rejoinder to bell and jones. *Social science & medicine*, 145:125–128, 2015.

[80] Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of Documentation*, 60(5):503–520, 2004.

[81] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pages 399–408. ACM, 2015.

[82] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*, 1:1–20, 2010.

[83] Donald B Rubin and Dorothy T Thayer. EM algorithms for ml factor analysis. *Psychometrika*, 47(1):69–76, 1982.

[84] Seyed Ehsan Saffari and Robiah Adnan. Zero-inflated Poisson regression models with right censored count data. *Matematika (Johor Bahru)*, 27(1):21–29, 2011.

[85] Nora Cate Schaeffer and Jennifer Dykema. Questions for surveys: current trends and future directions. *Public opinion quarterly*, 75(5):909–961, 2011.

[86] Nora Cate Schaeffer and Jennifer Dykema. Advances in the science of asking questions. *Annual Review of Sociology*, 46, 2020.

[87] Nora Cate Schaeffer and Stanley Presser. The science of asking questions. *Annual Review of Sociology*, 29(1):65–88, 2003.

[88] John Scott Long. *Regression models for categorical and limited dependent variables*, volume 7. Thousand Oaks: Sage Publications, 1997.

[89] Lei Shi. Distributed learning with indefinite kernels. *Analysis and Applications*, pages 1–29, 2019.

[90] Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, 26(2):153–172, 2007.

[91] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008.

[92] Ingo Steinwart, Don R. Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.

[93] Gilbert Strang. *Introduction to Linear Algebra, vol. 3.* Wellesley Cambridge Press, 1993.

[94] Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in Neural Information Processing Systems*, pages 1385–1392, 2005.

[95] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[96] Grace Wahba. *Spline models for observational data.* SIAM, Philadelphia, 1990.

[97] Cheng Wang and Ting Hu. Online minimum error entropy algorithm with unbounded sampling. *Anal. Appl. (Singap.)*, 17(2):293–322, 2019.

[98] Dan J Wang and Sarah A Soule. Social movement organizational collaboration: Networks of learning and the diffusion of protest tactics, 1960–1995. *American Journal of Sociology*, 117(6):1674–1722, 2012.

[99] Yang Yang, Sam Schulhofer-Wohl, Wenjiang J Fu, and Kenneth C Land. The intrinsic estimator for age-period-cohort analysis: what it is and how to use it. *American Journal of Sociology*, 113(6):1697–1736, 2008.

[100] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constr. Approx.*, 26(2):289–315, 2007.

[101] Achim Zeileis, Christian Kleiber, and Simon Jackman. Regression models for count data in R. *Journal of statistical software*, 27(8):1–25, 2008.

[102] Tong Zhang. Effective dimension and generalization of kernel learning. In *Advances in Neural Information Processing Systems*, pages 454–461, 2002.

[103] Yulong Zhao, Jun Fan, and Lei Shi. Learning rates for regularized least squares ranking algorithm. *Anal. Appl. (Singap.)*, 15(6):815–836, 2017.

[104] Mu Zhu and Ali Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51(2):918–930, 2006.

[105] Xiang Zhu and Matthew Stephens. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *Ann. Appl. Stat.*, 11(3):1561–1592, 2017.

[106] Yushu Zhu. Toward community engagement: Can the built environment help? grassroots participation and communal space in chinese urban communities. *Habitat International*, 46:44–53, 2015.

[107] Yushu Zhu, Werner Breitung, and Si-ming Li. The changing meaning of neighbourhood attachment in chinese commodity housing estates: Evidence from guangzhou. *Urban Studies*, 49(11):2439–2457, 2012.

[108] Yushu Zhu and Qiang Fu. Deciphering the civic virtue of communal space: Neighborhood attachment, social capital, and neighborhood participation in urban china. *Environment and Behavior*, 49(2):161–191, 2017.

[109] Christopher JW Zorn. An analytic and empirical examination of zero-inflated and hurdle poisson specifications. *Sociological Methods & Research*, 26(3):368–400, 1998.

[110] Laurent Zwald and Gilles Blanchard. On the convergence of eigenspaces in kernel principal component analysis. In *Advances in neural information processing systems*, pages 1649–1656, 2006.

[111] Laurent Zwald, Gilles Blanchard, Pascal Massart, and Régis Vert. Kernel projection machine: a new tool for pattern recognition. In *Advances in Neural Information Processing Systems*, pages 1649–1656, 2005.