Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Data Thinning for Poisson Factor Models and its Applications

Zhijing Wang, Peirong Xu, Hongyu Zhao & Tao Wang

To cite this article: Zhijing Wang, Peirong Xu, Hongyu Zhao & Tao Wang (20 Oct 2025): Data Thinning for Poisson Factor Models and its Applications, Journal of the American Statistical Association, DOI: 10.1080/01621459.2025.2546577

To link to this article: https://doi.org/10.1080/01621459.2025.2546577

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2025, VOL. 00, NO. 0, 1–14: Theory and Methods
https://doi.org/10.1080/01621459.2025.2546577

Data Thinning for Poisson Factor Models and its Applications

Zhijing Wang[a], Peirong Xu[a], Hongyu Zhao[b,c], and Tao Wang[a,c,d]

[a]Department of Statistics, Shanghai Jiao Tong University, Shanghai, China; [b]Department of Biostatistics, Yale University, New Haven, CT; [c]SJTU-Yale Joint Center of Biostatistics and Data Science, Shanghai Jiao Tong University, Shanghai, China; [d]Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

The Poisson factor model is a powerful tool for dimension reduction and visualization of large-scale count datasets across diverse domains. Despite the availability of efficient algorithms for estimating factors and loadings, existing methods either require prior knowledge of the number of factors, or resort to ad hoc criteria for its determination. This article proposes a novel data-driven criterion called Information Criterion via Data Thinning (ICDT), leveraging the thinning property of the Poisson distribution. Unlike traditional data splitting, data thinning partitions the count matrix into training and validation sets while preserving both the distribution and the underlying data structure. Interestingly, the validation error can be decomposed into the training error plus a covariance penalty. A simple estimator of the covariance penalty is obtained, leading to the development of ICDT. The selection consistency of ICDT is derived when both the sample size and the number of variables diverge to infinity. The proposed methodology is extended to dimension reduction in regression by incorporating the response inversely into the Poisson factor model. Extensive simulated examples and two real data applications are used to evaluate the performance of ICDT and compare it with existing criteria. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received July 2024
Accepted July 2025

KEYWORDS
Degrees of freedom; Factor analysis; Model selection; Poisson thinning; Supervised dimension reduction

1. Introduction

Count data, representing the frequency of events or occurrences, are prevalent across diverse fields such as life sciences, social sciences, finance, and marketing. With the expanding scope and scale of data collection, researchers are increasingly using count data to uncover interesting insights, reveal patterns and trends, and inform decision-making processes (Wang et al. 2018; Kelly, Manela, and Moreira 2021; Cohn, Liu, and Wardlaw 2022). However, in modern practice, count data typically involve a large number of observations on numerous variables, posing challenges in statistical modeling, computational efficiency, and interpretation.

Factor analysis, a versatile statistical method with applications across various disciplines (Bartholomew, Knott, and Moustaki 2011), explains patterns of correlations among observed variables by reducing the dimensionality of data, representing a large number of variables with a smaller number of factors. While various methods have been proposed for factor extraction and determining the number of factors, existing approaches have primarily focused on continuous data and linear factor models (e.g., Bai and Li 2012; Li et al. 2018; Fan, Guo, and Zheng 2022). However, these methods, typically based on a squared loss or the Gaussian distribution, may encounter computational and interpretational challenges when applied to count data. To account for the non-negativity and discreteness of observed counts, it is more suitable to model them using the Poisson or negative binomial distribution.

Some progress has been made in the framework of generalized factor models (GFMs, also known as generalized latent variable models), which extend linear factor models to non-Gaussian data (Chen and Li 2022; Kidziński et al. 2022; Wang 2022). To estimate parameters in GFMs, one can integrate the random factors out and maximize the marginal log-likelihood function. Approximating the marginal likelihood using penalized quasi-likelihood, Kidziński et al. (2022) introduced two efficient algorithms for fitting large-scale GFMs and proposed using nuclear norm regularization and cross-validation to select the number of latent variables or factors. Nevertheless, the theoretical properties of parameter estimation and model selection have not been established. On the other hand, treating the latent variables as fixed model parameters, Chen and Li (2022) introduced a joint-likelihood-based information criterion for determining the number of factors in GFMs, and developed an alternating optimization procedure for maximizing the joint likelihood. Under mild conditions, they proved the consistency of the proposed criterion when both the numbers of samples and variables grow to infinity. A similar approach was independently proposed by Liu et al. (2023) in a more general setting where the observed variables are of mixed types. A different line of research employs Bayesian methods, as explored in Bhattacharya and

CONTACT Tao Wang *neowangtao@sjtu.edu.cn* Department of Statistics, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

Dunson (2011), Srivastava, Engelhardt, and Dunson (2017), and Legramanti, Durante, and Dunson (2020).

In a seminal work, Bai and Ng (2002) suggested several information criteria for use in linear factor models, each employing an ad hoc penalty term. These criteria have since been adopted for generalized factor models. Let n, p, and k denote the numbers of samples, variables, and factors, respectively. Two candidate penalty terms take the form $k\left\{(n \vee p)/np\right\} \ln\left\{np/(n \vee p)\right\}$ and $k\left\{(n+p)/np\right\} \ln\left\{np/(n+p)\right\}$. The former was recommended by Chen and Li (2022), while the latter was adopted by Liu et al. (2023), and both were considered by Jentsch, Lee, and Mammen (2021). However, despite their widespread application, there is currently no theoretical guidance on how to differentiate or choose among them. In linear factor models, an alternative way of consistently selecting the number of factors is the ratio-based rule (Lam, Yao, and Bathia 2011; Ahn and Horenstein 2013). The extension of such a rule to generalized factor models was recently examined by Jentsch, Lee, and Mammen (2021), but without a theoretical guarantee.

In this article, we consider the Poisson factor model for extracting meaningful insights from large-scale count datasets, and propose two data-driven criteria for model selection by leveraging the thinning property of the Poisson distribution (Last and Penrose 2017). Loosely speaking, the technique of data thinning decomposes a random variable into two independent variables, each of which, up to a known parameter scaling, follows the same distribution as the original variable (refer to Leiner et al. 2023, Neufeld et al. 2024, and the references therein for further details).

The first criterion, Validation Set via Data Thinning (VSDT), involves decomposing the observed count matrix into two matrices, fitting a Poisson factor model to one matrix, and evaluating the fitted model using the other matrix. However, theoretical analysis shows that VSDT has a tendency to overestimate the true number of factors under the assumption of its existence. As a by-product of VSDT, an unbiased estimator of the covariance penalty for the Poisson factor model is obtained, which, to our knowledge, represents a novel contribution to the factor analysis literature. The asymptotic property of the estimator is derived under mild conditions. This leads directly to the second criterion, Information Criterion via Data Thinning (ICDT), which takes the form of a generalized information criterion. The selection consistency of ICDT is established as both the sample size and the number of variables grow to infinity.

Factor analysis serves as a technique for extracting a few factors that explain a large proportion of the variability in the data. Typically, these factors are treated as predictors and are connected to a response variable in a forward regression. However, factor analysis is an unsupervised technique because the response is not used. Model-based inverse regression provides a framework of dimension reduction in regression, wherein the conditional distribution of the predictors given the response is modeled to construct low-dimensional substitutes for the predictors (Cook and Forzani 2008). Although some progress has been made for count-valued predictors (Taddy 2013, 2015; Pang, Zhao, and Wang 2024), the theoretical properties of these methods are either lacking or limited to the setting where the number of predictors diverges slowly as the sample size increases. To make the factors relevant for predicting the response, the Poisson factor inverse regression model is introduced, which inversely incorporates the response into the Poisson factor model. The proposed methodology is then extended to the inverse regression model. Theoretical properties of the estimates, as well as the selection consistency of ICDT, are derived under more relaxed assumptions on the relationship between the number of predictors and the sample size.

2. Poisson Factor Model for Multivariate Count Data

Let $\boldsymbol{X} = (x_{ij})$ denote an $n \times p$ matrix of observed counts, where rows $i = 1, \ldots, n$ are samples and columns $j = 1, \ldots, p$ are variables or features. The Poisson factor model assumes that, conditionally on a vector of $q < p$ latent variables or factors, $\boldsymbol{f}_i$, all components of the ith row, $X_{i1}, \ldots, X_{ip}$, are independent, and each follows a Poisson distribution,

$$\begin{cases} X_{ij} \sim \text{Poisson}\left(\lambda_{ij}\right), \\ \lambda_{ij} = \exp\left(\boldsymbol{l}_j^\top \boldsymbol{f}_i + \alpha_j\right), \end{cases} \tag{1}$$

where α_j denotes a feature-specific intercept and $\boldsymbol{l}_j$ is a q-vector of factor loadings.

As mentioned previously, the Poisson factor model, widely used across diverse domains, is a prominent tool for extracting meaningful insights from multivariate count data. For example, the factors $\boldsymbol{f}_i$ are not only useful for data visualization but can also serve as inputs for downstream tasks, such as clustering and supervised learning (Kidziński et al. 2022; Zhang, Xu, and Zhu 2022).

Let $\boldsymbol{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n)^\top \in \mathbb{R}^{n \times q}$ and $\boldsymbol{L} = (\boldsymbol{l}_1, \ldots, \boldsymbol{l}_p)^\top \in \mathbb{R}^{p \times q}$. Neither $\boldsymbol{L}$ nor $\boldsymbol{F}$ is identifiable. Indeed, for any $q \times q$ nonsingular matrix $\boldsymbol{V}$, model (1) remains unchanged if we rotate $\boldsymbol{F}$ and $\boldsymbol{L}^\top$ to $\boldsymbol{FV}$ and $\boldsymbol{V}^{-1}\boldsymbol{L}^\top$, respectively. To avoid rotation invariance and ensure parameter identifiability, we impose the following assumptions:

(A1) $n^{-1}\sum_{i=1}^{n} \boldsymbol{f}_i = \boldsymbol{0}_q$ and $n^{-1}\boldsymbol{F}^\top\boldsymbol{F} = \boldsymbol{I}_q$, where $\boldsymbol{0}_q$ is the q-vector of zeros, and $\boldsymbol{I}_q$ is the $q \times q$ identity matrix;

(A2) $\boldsymbol{L}^\top\boldsymbol{L}$ is a diagonal matrix with decreasing diagonal elements;

(A3) The element with maximum absolute value in each column of $\boldsymbol{L}$ is positive.

Bai and Li (2012) summarized several sets of identifiability conditions for linear factor models. These conditions have been adopted or modified for generalized factor models (Jentsch, Lee, and Mammen 2021; Chen and Li 2022; Liu et al. 2023). Therefore, the assumptions (A1)–(A3) are just one set of conditions that achieve identifiability for Poisson factor models. In the latent variable modeling literature, where the factors are random, it is common to assume that the vector of factors has a zero mean vector and an identity covariance matrix. Assumption (A1), which imposes a set of constraints on factors, is a sample counterpart. Given (A1), assumptions (A2) and (A3) on loadings are natural, as the factor matrix $\boldsymbol{F}$ and the loading matrix $\boldsymbol{L}$ can be uniquely determined up to a diagonal matrix whose entries are 1 or -1 on the diagonals. Specifically, (A2) accommodates varying signal strengths across different factors, while (A3) imposes a sign constraint on one element in each

column of the loading matrix. Note that in this article, we assume that the factors are nonrandom. If they were assumed to be random, then our method could be interpreted as conditioning on them. See the supplementary material for a proof of the following proposition.

Proposition 1. Let $(\boldsymbol{F}, \boldsymbol{L}, \boldsymbol{\alpha})$ and $(\boldsymbol{F}', \boldsymbol{L}', \boldsymbol{\alpha}')$ be two set of parameters satisfying model (1) and Assumptions (A1)–(A3), then we have $\boldsymbol{F} = \boldsymbol{F}', \boldsymbol{L} = \boldsymbol{L}'$, and $\boldsymbol{\alpha} = \boldsymbol{\alpha}'$.

We adopt a joint likelihood approach for fitting the Poisson factor model, due to the availability of efficient algorithms that can scale to large datasets, as well as the established theoretical results applicable in high-dimensional settings where both n and p diverge to infinity (Chen and Li 2022; Kidziński et al. 2022; Liu et al. 2023). Let $\boldsymbol{\theta}_j = (\alpha_j, \boldsymbol{l}_j^\top)^\top \in \mathbb{R}^{q+1}$ and $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)^\top \in \mathbb{R}^{p\times(q+1)}$. The joint log-likelihood of the observed data is

$$\begin{aligned} l(\boldsymbol{F}, \boldsymbol{\Theta}; \boldsymbol{X}) &= \sum_{i=1}^{n}\sum_{j=1}^{p} l\left(\boldsymbol{f}_i, \boldsymbol{\theta}_j; x_{ij}\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{p}\Big[\left\{x_{ij}\left(\boldsymbol{l}_j^\top \boldsymbol{f}_i + \alpha_j\right) - \exp\left(\boldsymbol{l}_j^\top \boldsymbol{f}_i + \alpha_j\right)\right\} \\ &\quad - \ln(x_{ij}!)\Big]. \end{aligned}$$

We maximize the joint log-likelihood via an alternating two-step procedure. In Step 1, conditional on $\boldsymbol{F}^{(t)}$, we update $\boldsymbol{\Theta}$ by solving

$$\boldsymbol{\theta}_j^{(t+1)} = \arg\max_{\boldsymbol{\theta}_j} \sum_{i=1}^{n} l\left(\boldsymbol{f}_i^{(t)}, \boldsymbol{\theta}_j; x_{ij}\right),$$

for $j = 1, \ldots, p$, and then we adjust $\boldsymbol{\Theta}^{(t+1)}$ to meet Assumptions (A2)–(A3). In Step 2, conditional on $\boldsymbol{\Theta}^{(t+1)}$, we update $\boldsymbol{F}$ by solving

$$\boldsymbol{f}_i^{(t+1)} = \arg\max_{\boldsymbol{f}_i} \sum_{j=1}^{p} l\left(\boldsymbol{f}_i, \boldsymbol{\theta}_j^{(t+1)}; x_{ij}\right),$$

for $i = 1, \ldots, n$, and then we adjust $\boldsymbol{F}^{(t+1)}$ to fulfill Assumption (A1). In our algorithm, we iterate through Step 1 and Step 2 until the change in log-likelihood relative to the updated log-likelihood becomes sufficiently small. Additional implementation details, including the pseudo code, can be found in the supplementary material.

3. Determining the Number of Factors

In this section, we introduce two novel criteria for determining the number of factors, both leveraging the data thinning technique (Last and Penrose 2017; Neufeld et al. 2024).

3.1. Validation Set via Data Thinning

We begin this subsection by presenting the thinning property of the Poisson distribution.

Proposition 2 (Theorem 5.8 in Last and Penrose 2017). If $X \sim$ Poisson (λ) and $Z \mid X \sim$ Binomial(X, π), where $0 < \pi < 1$, then $Z \sim$ Poisson $(\pi\lambda)$ and $X - Z \sim$ Poisson $((1-\pi)\lambda)$. Furthermore, Z and $X - Z$ are independent.

The thinning property is not limited to the Poisson distribution. It also holds for the normal, multinomial, and gamma distributions, and more generally, for convolution-closed distributions (Neufeld et al. 2024). More recently, the concept of data thinning has been extended to natural and general exponential families, as well as to distributions beyond both the convolution-closed and exponential families; see Dharamshi et al. (2025) for more details. According to Proposition 2, for $\pi = 0.5$, the observed data matrix $\boldsymbol{X}$ can be thinned into two independent and identically distributed parts, $\boldsymbol{X}^A$ and $\boldsymbol{X}^B$. Consequently, for a candidate set of models, we can perform model assessment and selection by applying each model to $\boldsymbol{X}^A$, and then validate it using $\boldsymbol{X}^B$. This leads to a Validation Set via Data Thinning (VSDT) criterion for choosing the number of factors.

For a Poisson factor model containing k factors, denote by $\hat{\boldsymbol{F}}^{A,k}$ and $\hat{\boldsymbol{\Theta}}^{A,k}$ the factor matrix and loading matrix obtained by fitting the model to $\boldsymbol{X}^A$. We measure the quality of fit of the model on $\boldsymbol{X}^B$ by evaluating the log-likelihood of $\boldsymbol{X}^B$ at $\hat{\boldsymbol{F}}^{A,k}$ and $\hat{\boldsymbol{\Theta}}^{A,k}$:

$$\begin{aligned} l\left(\hat{\boldsymbol{F}}^{A,k}, \hat{\boldsymbol{\Theta}}^{A,k}; \boldsymbol{X}^B\right) = \sum_{i=1}^{n}\sum_{j=1}^{p}\Big\{x_{ij}^B\left(\hat{\boldsymbol{l}}_j^{A,k\top}\hat{\boldsymbol{f}}_i^{A,k} + \hat{\alpha}_j^{A,k}\right) \\ -\frac{1}{2}\exp\left(\hat{\boldsymbol{l}}_j^{A,k\top}\hat{\boldsymbol{f}}_i^{A,k} + \hat{\alpha}_j^{A,k}\right)\Big\}, \end{aligned}$$

where terms independent of k have been omitted. The VSDT criterion is given by

$$\begin{aligned} \hat{q}_1 &= \operatorname{argmax}_k l\left(\hat{\boldsymbol{F}}^{A,k}, \hat{\boldsymbol{\Theta}}^{A,k}; \boldsymbol{X}^B\right) \\ &= \operatorname{argmax}_k \sum_{i=1}^{n}\sum_{j=1}^{p}\Big\{x_{ij}^B\left(\hat{\boldsymbol{l}}_j^{A,k\top}\hat{\boldsymbol{f}}_i^{A,k} + \hat{\alpha}_j^{A,k}\right) \\ &\quad -\frac{1}{2}\exp\left(\hat{\boldsymbol{l}}_j^{A,k\top}\hat{\boldsymbol{f}}_i^{A,k} + \hat{\alpha}_j^{A,k}\right)\Big\}. \end{aligned}$$

Data thinning bears similarity to data splitting, as both are resampling methods rooted in the concept of randomization. While data splitting has gained considerable popularity and can be applied to various statistical problems (Rinaldo, Wasserman, and Gsell 2019; Dai et al. 2023), its utility within the framework of factor analysis is limited due to the disruption of the data structure when dividing either the samples or features into two (or more) parts (Owen and Wang 2016). In contrast, data thinning offers an advantage in this regard by partitioning the data matrix into two (or more) parts while preserving not only the distribution but also the underlying data structure.

Theorem 1. Assume that Conditions (C1)–(C3) in the supplementary material hold. If $n = o(p^4)$ and $\ln p = o(n)$, then we have $\mathbb{P}(\hat{q}_1 \geq q) \to 1$, as $n, p \to \infty$.

The VSDT criterion evaluates a candidate set of models based on their predictive ability, similar in spirit to cross-validation

(Hastie et al. 2009). It is known that traditional cross-validation, which is based on sample splitting, tends to overfit the model (Shao 1993; Jin, Miao, and Su 2021). Similarly, VSDT can be susceptible to over-selection; see Section 5 for details.

It is generally a good idea to set a random seed when performing a resampling procedure that contains an element of randomness, so that the results obtained can be reproduced precisely at a later time. This is the case for the VSDT criterion. Alternatively, we can repeat VSDT a couple of times, and then aggregate the results by taking a majority vote.

3.2. Covariance Penalty

The following proposition regarding VSDT is both simple and powerful, offering an insightful characteristic of this criterion.

Proposition 3. Let $\mathbb{E}(\cdot)$ indicate expectation over both $\boldsymbol{X}^A$ and $\boldsymbol{X}^B$. Then

$$\begin{aligned}\mathbb{E}\left\{-l\left(\hat{\boldsymbol{F}}^{A,k}, \hat{\boldsymbol{\Theta}}^{A,k}; \boldsymbol{X}^B\right)\right\} &= \mathbb{E}\left\{-l\left(\hat{\boldsymbol{F}}^{A,k}, \hat{\boldsymbol{\Theta}}^{A,k}; \boldsymbol{X}^A\right)\right\} \\ &\quad + \sum_{i=1}^{n}\sum_{j=1}^{p} \mathrm{Cov}\left(x_{ij}^A,\ \hat{\boldsymbol{l}}_j^{A,k\top}\hat{\boldsymbol{f}}_i^{A,k} + \hat{\alpha}_j^{A,k}\right).\end{aligned}$$

If we treat $\boldsymbol{X}^A$ as the training dataset and $\boldsymbol{X}^B$ as the validation dataset, then, on average, the first term on the right-hand side is the training error, while the second term takes the form of a covariance penalty (Efron and Hastie 2016). Proposition 3 then says that the training error underestimates the true error by the covariance penalty.

The covariance penalty is a fundamental concept in statistics, closely linked to the notion of degrees of freedom (Efron and Hastie 2016). However, despite their importance, there is a lack of available results regarding either covariance penalty or degrees of freedom in the context of the Poisson factor model. In general, the form of the covariance penalty depends on the context assumed for the prediction problem. While it can sometimes be determined analytically, more often, it necessitates computationally intensive methods such as the bootstrap (Efron and Hastie 2016). Let

$$\mathrm{df}_k = \sum_{i=1}^{n}\sum_{j=1}^{p} \mathrm{cov}\left(x_{ij}^A,\ \hat{\boldsymbol{l}}_j^{A,k\top}\hat{\boldsymbol{f}}_i^{A,k} + \hat{\alpha}_j^{A,k}\right),$$

and

$$\hat{\mathrm{df}}_k = l\left(\hat{\boldsymbol{F}}^{A,k}, \hat{\boldsymbol{\Theta}}^{A,k}; \boldsymbol{X}^A\right) - l\left(\hat{\boldsymbol{F}}^{A,k}, \hat{\boldsymbol{\Theta}}^{A,k}; \boldsymbol{X}^B\right).$$

Somewhat miraculously, Proposition 3 shows that, for the Poisson factor model, $\hat{\mathrm{df}}_k$ is an unbiased estimate of df_k. The following proposition analyzes $\hat{\mathrm{df}}_k$ asymptotically.

Proposition 4. Assume that Conditions (C1)–(C3) in the supplementary material hold. For $k \geq q$, omitting terms that are independent of k, we have

$$\hat{\mathrm{df}}_k = (k+1)p + kn + o_p(n+p).$$

As previously mentioned, data thinning involves an element of randomness. An improved stable estimate of df_k can be obtained by repeating thinning $M > 1$ times and then aggregating the results by calculating the mean. Let $\boldsymbol{X}_m^A$ and $\boldsymbol{X}_m^B$ denote the data matrices from the mth repetition, for $m = 1, \ldots, M$. We propose to estimate df_k by

$$\frac{1}{M}\sum_{m=1}^{M}\left\{l\left(\hat{\boldsymbol{F}}_m^{A,k}, \hat{\boldsymbol{\Theta}}_m^{A,k}; \boldsymbol{X}_m^A\right) - l\left(\hat{\boldsymbol{F}}_m^{A,k}, \hat{\boldsymbol{\Theta}}_m^{A,k}; \boldsymbol{X}_m^B\right)\right\}.$$

3.3. Information Criterion via Data Thinning

A criterion for selecting the number of factors is considered consistent when, in the scenario where the true model is included among the candidate models, the probability of selecting the true model approaches one. According to Theorem 1, the VSDT criterion may not be consistent without additional stringent conditions. Several information criteria have been proposed for generalized factor models, and the consistency of some is established as both the sample size (n) and the number of variables (p) grow to infinity (Jentsch, Lee, and Mammen 2021; Chen and Li 2022; Liu et al. 2023). However, these criteria often employ an ad hoc penalty term without theoretical support and interpretation.

To address this issue, we introduce an Information Criterion via Data thinning (ICDT) by leveraging the unbiasedness of $\hat{\mathrm{df}}_k$ as an estimate for df_k. Specifically, the criterion has the form

$$\hat{q}_2 = \arg\max_k \left\{l\left(\hat{\boldsymbol{F}}^k, \hat{\boldsymbol{\Theta}}^k; \boldsymbol{X}\right) - \hat{\mathrm{df}}_k \times f(n,p)\right\},$$

where $f(n,p) > 0$ is a penalty function, and needs to be specified.

The following theorem asserts the selection consistency of ICDT.

Theorem 2. Assume that Conditions (C1)–(C3) in the supplementary material hold. If $n = o(p^4)$, $\ln p = o\left(n^{3/4}\right)$, $f(n,p) = O\left(\min(\sqrt{n}, \sqrt{p})\right)$, and $f^{-1}(n,p) = o(1)$, then we have $\mathbb{P}(\hat{q}_2 = q) \to 1$, s $n, p \to \infty$.

The condition $f^{-1}(n,p) = o(1)$, or equivalently $f(n,p) \to \infty$, ensures that when the working number of factors exceeds the true number, the penalty term dominates and the probability of overestimating the number of factors converges to zero. Besides, the condition $f(n,p) = O\left(\min(\sqrt{n}, \sqrt{p})\right)$ guarantees that when the working number of factors is smaller than the true number, the difference in the likelihood function becomes the dominant term, ensuring that the probability of underestimating the number of factors tends to zero. These two conditions on the penalty function $f(n,p)$ are mild and commonly assumed to ensure the consistency of information criteria for selecting the number of factors (Bai and Ng 2002; Liu et al. 2023). We take $f(n,p) = \ln\left\{np/(n+p)\right\}$ in our numerical studies.

4. Poisson Factor Augmented Inverse Regression

In this section, we consider the regression or classification setting where, alongside the high-dimensional count vector $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$, there is a response variable of interest, denoted by

Y. We are interested in understanding the relationship between $\boldsymbol{X}$ and Y. The Poisson factor model (1), along with the proposed criteria, serves as a method for extracting a few factors that explain a large proportion of the variability in the count vector $\boldsymbol{X}$. However, the factors are identified in an unsupervised manner, as the response Y is not used to determine them. Consequently, the Poisson factor model suffers from a limitation: there is no guarantee that the extracted factors will be relevant for predicting the response. One way of addressing this issue is to incorporate the response inversely into the Poisson factor model, leading to the Poisson factor inverse regression model:

$$\begin{cases} X_{ij} \mid Y_i = y_i \sim \text{Poisson}\left(\lambda_{ij}\right), \\ \lambda_{ij} = \exp\left(\boldsymbol{l}_j^\top \boldsymbol{f}_i + \boldsymbol{b}_j^\top \boldsymbol{s}_{y_i} + \alpha_j\right), \end{cases} \tag{2}$$

where $\boldsymbol{s}_y \in \mathbb{R}^d$ is a known function of the response and $\boldsymbol{b}_j \in \mathbb{R}^d$ is the corresponding coefficient vector.

Similar to the Poisson factor model, the inverse regression model is used to identify factors that explain the count vector. However, in this model, the response variable supervises the identification of these factors, resulting in factors that explain both the count vector and the response. Let $\boldsymbol{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p)^\top \in \mathbb{R}^{p\times d}$. This claim is made precise in the following proposition.

Proposition 5. Under model (2), $Y \perp\!\!\!\perp \boldsymbol{X} \mid (\boldsymbol{L}, \boldsymbol{B})^\top \boldsymbol{X}$. Furthermore, $(\boldsymbol{L}, \boldsymbol{B})^\top \boldsymbol{X}$ is minimal sufficient for the regression of Y on $\boldsymbol{X}$.

In scenarios with high-dimensional predictors, conducting forward regressions can be challenging, especially in the absence of a pre-specified model. In contrast, fitting an inverse regression model of the predictors on the response is more manageable and facilitates the construction of low-dimensional summaries of the predictors relevant to the forward regression. In particular, for multivariate counts, the multinomial inverse regression serves as a tool for supervised dimension reduction (Taddy 2013; Pang, Zhao, and Wang 2024). Given that the multinomial distribution can be approximated by a product of independent Poisson distributions (Taddy 2015), model (2) can be regarded as an extension of the multinomial inverse regression model.

While classical sufficient dimension reduction methods also employ the idea of inverse regression, they are largely moment-based, making them unsuitable for count predictors (Li 2018). Moreover, in contrast to the proposed method, they are typically designed for low-dimensional settings and often rely on sparsity assumptions when extended to high-dimensional data (Li 2018).

4.1. Estimation

From the perspective of dimension reduction in regression, only the subspace spanned by the columns of $(\boldsymbol{L}, \boldsymbol{B})$ is identifiable. Such a space is called a dimension-reduction subspace (Cook and Forzani 2008). Nevertheless, as with model (1), the parameters in model (2) can still be identified under certain assumptions.

Assumption (A1') $n^{-1}\sum_{i=1}^n \boldsymbol{f}_i = \boldsymbol{0}_q$, $n^{-1}\boldsymbol{F}^\top\boldsymbol{F} = \boldsymbol{I}_q$, and $\sum_{i=1}^n \boldsymbol{s}_{y_i} f_{il} = \boldsymbol{0}_d$ for $1 \le l \le q$.

Proposition 6. Let $(\boldsymbol{F}, \boldsymbol{L}, \boldsymbol{B}, \boldsymbol{\alpha})$ and $\left(\boldsymbol{F}', \boldsymbol{L}', \boldsymbol{B}', \boldsymbol{\alpha}'\right)$ be two sets of parameters satisfying model (2). Then, under Assumptions (A1'), (A2), and (A3), we have $\boldsymbol{F} = \boldsymbol{F}', \boldsymbol{L} = \boldsymbol{L}', \boldsymbol{B} = \boldsymbol{B}'$, and $\boldsymbol{\alpha} = \boldsymbol{\alpha}'$.

As before, we use a joint likelihood approach to fitting the inverse regression model. Let $\boldsymbol{\phi}_j = \left(\alpha_j, \boldsymbol{l}_j^\top, \boldsymbol{b}_j^\top\right)^\top$ and $\boldsymbol{\Phi} = \left(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_p\right)^\top$. The joint log-likelihood function is

$$\begin{aligned} l\left(\boldsymbol{F}, \boldsymbol{\Phi}; \boldsymbol{X}, \boldsymbol{y}\right) &= \sum_{i=1}^n \sum_{j=1}^p l\left(\boldsymbol{f}_i, \boldsymbol{\phi}_j; x_{ij}, y_i\right) \\ &= \sum_{i=1}^n \sum_{j=1}^p \Big[\Big\{x_{ij}\left(\boldsymbol{l}_j^\top \boldsymbol{f}_i + \boldsymbol{b}_j^\top \boldsymbol{s}_{y_i} + \alpha_j\right) \\ &\quad - \exp\left(\boldsymbol{l}_j^\top \boldsymbol{f}_i + \boldsymbol{b}_j^\top \boldsymbol{s}_{y_i} + \alpha_j\right)\Big\}\Big], \end{aligned}$$

where terms independent of $\boldsymbol{F}$ and $\boldsymbol{\Phi}$ have been omitted.

Let $\boldsymbol{F}^*$ and $\boldsymbol{\Phi}^*$ denote the true values of $\boldsymbol{F}$ and $\boldsymbol{\Phi}$, respectively. To obtain the maximum joint likelihood estimates of $\boldsymbol{F}^*$ and $\boldsymbol{\Phi}^*$, we develop an alternative maximization algorithm; see the supplementary material for further details. We establish the asymptotic behavior of $\hat{\boldsymbol{F}}$ and $\hat{\boldsymbol{\Phi}}$ in the following theorem.

Theorem 3. Assume that Conditions (C1')–(C3') in the supplementary material hold. Then we have

$$\begin{aligned} \left\|\hat{\boldsymbol{f}}_i - \boldsymbol{f}_i^*\right\|_2 &= O_p\left(\frac{1}{C_{np}}\right), \\ \sup_i \left\|\hat{\boldsymbol{f}}_i - \boldsymbol{f}_i^*\right\|_2^2 &= O_p\left(\frac{1}{n} + \frac{n^{1/4}}{p}\right), \\ \left\|\hat{\boldsymbol{\phi}}_j - \boldsymbol{\phi}_j^*\right\|_2 &= O_p\left(\frac{1}{C_{np}}\right), \\ \sup_j \left\|\hat{\boldsymbol{\phi}}_j - \boldsymbol{\phi}_j^*\right\|_2^2 &= O_p\left(\frac{1}{p} + \frac{\ln p}{n}\right), \end{aligned}$$

where $C_{np} = \min\{\sqrt{n}, \sqrt{p}\}$.

Interestingly, according to Theorem 3, there is a "blessing of dimensionality," that is, the larger the p, the more precise the estimates of factors and loadings. Under stronger conditions, such as the boundedness of factors, one can show that

$$\begin{aligned} \frac{1}{\sqrt{n}}\left\|\hat{\boldsymbol{F}} - \boldsymbol{F}^*\right\|_F &= O_p\left(\frac{1}{C_{np}}\right), \\ \frac{1}{\sqrt{p}}\left\|\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}^*\right\|_F &= O_p\left(\frac{1}{C_{np}}\right). \end{aligned}$$

See, for example, Chen and Li (2022).

4.2. Determining the Number of Sufficient Predictors

Just like the factors in an unsupervised factor model, the sufficient reduction scores, $(\boldsymbol{L}, \boldsymbol{B})^\top \boldsymbol{X}$, in the inverse regression model (2), are valuable for visualization and can be used as inputs in a forward regression. Consequently, the selection of the structural dimension, $q + d$, becomes crucial. In certain scenarios, there may exist a natural choice for $\boldsymbol{s}_y$, and hence d. In this article, for regression tasks, we simply take $\boldsymbol{s}_y = y$, resulting

in $d = 1$, while for classifications involving g categories, we set $\boldsymbol{s}_y$ to be the vector of indicator functions, yielding $d = g$. It remains to determine the number of factors, q.

Under a supervised learning framework, there are many methods for choosing the appropriate value of q Cook and Forzani (2008). In this article, we adapt the ICDT method to the inverse regression model. Specifically, let

$$\mathrm{df}_k = \sum_{i=1}^{n}\sum_{j=1}^{p} \mathrm{Cov}\left(x_{ij}^A,\ \hat{\boldsymbol{l}}_j^{A,k\top}\hat{\boldsymbol{f}}_i^{A,k} + \hat{\boldsymbol{b}}_j^{A,k\top}\boldsymbol{s}_{y_i} + \hat{\alpha}_j^{A,k}\right)$$

and

$$\hat{\mathrm{df}}_k = l\left(\hat{\boldsymbol{F}}^{A,k}, \hat{\boldsymbol{\Phi}}^{A,k}; \boldsymbol{X}^A, \boldsymbol{y}\right) - l\left(\hat{\boldsymbol{F}}^{A,k}, \hat{\boldsymbol{\Phi}}^{A,k}; \boldsymbol{X}^B, \boldsymbol{y}\right).$$

We consider the criterion

$$\hat{q} = \arg\max_k \left\{ l\left(\hat{\boldsymbol{F}}^k, \hat{\boldsymbol{\Phi}}^k; \boldsymbol{X}, \boldsymbol{y}\right) - \hat{\mathrm{df}}_k \times f(n,p) \right\}.$$

Theorem 4. Assume that Conditions (C1')–(C3') in the supplementary material hold. If $n = o(p^4)$, $\ln p = o\left(n^{3/4}\right)$, $f(n,p) = O\left(\min(\sqrt{n}, \sqrt{p})\right)$, and $f^{-1}(n,p) = o(1)$, then we have $\mathbb{P}(\hat{q} = q) \to 1$, as $n, p \to \infty$.

5. Simulation Studies

In this section, we use simulated examples to assess the finite sample performance of VSDT and ICDT in choosing the number of factors. To keep the space concise, we focus on a simulation study for the Poisson factor model and defer the simulations for the Poisson factor inverse regression model to the supplementary material. Both VSDT and ICDT were based on running data thinning 10 times. We compared ICDT with five alternative information criteria proposed by Bai and Ng (2002), labeled IC_1 through IC_5. Unlike ICDT, which is data-driven, these criteria are ad hoc and use the following penalty terms:

$$P_1 = k\ln(n)/n,\quad P_2 = k\ln(p)/p,\quad P_3 = k(n+p-k)\ln(np)/np,$$
$$P_4 = k\left\{(n\vee p)/np\right\}\ln\left\{(np)/(n\vee p)\right\},$$
$$P_5 = k\left\{(n+p)/np\right\}\ln\left\{(np)/(n+p)\right\}.$$

Chen and Li (2022) recommended P_4, while Liu et al. (2023) preferred P_5. We selected the true number of factors from the candidate set $\{1,\ldots,8\}$ for each information criterion.

While traditional cross-validation based on sample splitting cannot be directly applied to factor models, several variants have been proposed. Three of these are included for comparison: (a) CV_1: bi-cross-validation (Owen and Wang 2016), which involves holding out some rows and columns of the data matrix, fitting a factor model to the held-in data, and comparing the held-out data to the corresponding fitted values in terms of

Table 1. Performance of factor selection based on 100 data repetitions for various criteria in Example 1.

		IC_1	IC_2	IC_3	IC_4	IC_5	RR	CV_1	CV_2	CV_3	adaGibbs	VSDT	ICDT
n	p	Frequency of correct (over|under) selection											
100	100	93 (7|0)	93 (7|0)	82 (0|18)	93 (7|0)	97 (3|0)	26 (12|62)	0 (100|0)	0 (100|0)	64 (36|0)	54 (46|0)	95 (5|0)	99 (1|0)
	200	93 (7|0)	90 (10|0)	98 (1|1)	93 (7|0)	94 (6|0)	21 (30|49)	0 (100|0)	0 (100|0)	66 (34|0)	52 (48|0)	94 (6|0)	97 (3|0)
	500	92 (8|0)	83 (17|0)	98 (2|0)	92 (8|0)	93 (7|0)	3 (39|58)	0 (100|0)	0 (100|0)	75 (25|0)	58 (42|0)	95 (5|0)	97 (3|0)
	2000	90 (10|0)	0 (100|0)	95 (5|0)	90 (10|0)	90 (10|0)	3 (28|69)	0 (100|0)	0 (100|0)	64 (36|0)	/	87 (13|0)	93 (7|0)
200	100	93 (7|0)	96 (4|0)	99 (0|1)	96 (4|0)	99 (1|0)	61 (21|18)	0 (100|0)	0 (100|0)	75 (25|0)	66 (34|0)	97 (3|0)	99 (1|0)
	200	94 (6|0)	94 (6|0)	99 (1|0)	94 (6|0)	96 (4|0)	45 (29|26)	0 (100|0)	0 (100|0)	69 (31|0)	62 (38|0)	96 (4|0)	98 (2|0)
	500	92 (8|0)	88 (12|0)	96 (4|0)	92 (8|0)	93 (7|0)	26 (40|34)	0 (100|0)	0 (100|0)	63 (37|0)	63 (37|0)	90 (10|0)	94 (6|0)
	2000	95 (5|0)	14 (86|0)	99 (1|0)	95 (5|0)	97 (3|0)	14 (57|29)	0 (100|0)	0 (100|0)	61 (39|0)	/	95 (5|0)	97 (3|0)
500	100	90 (10|0)	99 (1|0)	100 (0|0)	99 (1|0)	99 (1|0)	81 (10|9)	0 (100|0)	0 (100|0)	75 (25|0)	55 (45|0)	98 (2|0)	99 (1|0)
	200	92 (8|0)	98 (2|0)	100 (0|0)	98 (2|0)	98 (2|0)	58 (23|19)	0 (100|0)	0 (100|0)	70 (30|0)	64 (36|0)	96 (4|0)	100 (0|0)
	500	96 (4|0)	96 (4|0)	100 (0|0)	96 (4|0)	96 (4|0)	70 (26|4)	0 (100|0)	0 (100|0)	56 (44|0)	58 (42|0)	94 (6|0)	100 (0|0)
	2000	94 (6|0)	86 (14|0)	98 (2|0)	94 (6|0)	95 (5|0)	52 (48|0)	0 (100|0)	0 (100|0)	41 (59|0)	/	90 (10|0)	99 (1|0)
2000	100	0 (100|0)	100 (0|0)	100 (0|0)	100 (0|0)	100 (0|0)	89 (10|1)	0 (100|0)	0 (100|0)	63 (37|0)	50 (50|0)	100 (0|0)	100 (0|0)
	200	10 (90|0)	99 (1|0)	99 (1|0)	99 (1|0)	99 (1|0)	90 (8|2)	0 (100|0)	0 (100|0)	31 (69|0)	48 (52|0)	98 (2|0)	100 (0|0)
	500	77 (23|0)	93 (7|0)	96 (4|0)	93 (7|0)	93 (7|0)	98 (2|0)	0 (100|0)	0 (100|0)	11 (89|0)	50 (50|0)	88 (12|0)	95 (5|0)
	2000	83 (17|0)	83 (17|0)	98 (2|0)	83 (17|0)	90 (10|0)	94 (3|3)	0 (100|0)	0 (100|0)	8 (92|0)	/	77 (23|0)	93 (7|0)
n	p	Mean of the number of over|under selected factors											
100	100	1.00|0.00	1.00|0.00	0.00|1.00	1.00|0.00	1.00|0.00	1.17|3.81	1.26|0.00	1.00|0.00	1.14|0.00	1.00|0.00	1.00|0.00	1.00|0.00
	200	1.00|0.00	1.00|0.00	1.00|1.00	1.00|0.00	1.00|0.00	1.50|3.71	1.59|0.00	1.03|0.00	1.12|0.00	1.00|0.00	1.00|0.00	1.00|0.00
	500	1.12|0.00	1.12|0.00	1.00|0.00	1.12|0.00	1.14|0.00	1.33|3.76	2.82|0.00	1.27|0.00	1.08|0.00	1.00|0.00	1.20|0.00	1.00|0.00
	2000	1.10|0.00	2.83|0.00	1.00|0.00	1.10|0.00	1.10|0.00	1.75|3.62	3.00|0.00	2.25|0.00	1.17|0.00	/	1.15|0.00	1.00|0.00
200	100	1.00|0.00	1.00|0.00	0.00|1.00	1.00|0.00	1.00|0.00	1.24|3.78	1.52|0.00	1.02|0.00	1.16|0.00	1.00|0.00	1.00|0.00	1.00|0.00
	200	1.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.17|4.00	2.63|0.00	1.44|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00
	500	1.25|0.00	1.33|0.00	1.00|0.00	1.25|0.00	1.00|0.00	1.35|4.00	3.00|0.00	2.86|0.00	1.05|0.00	1.00|0.00	1.00|0.00	1.00|0.00
	2000	1.00|0.00	1.81|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.61|3.66	3.00|0.00	3.00|0.00	1.08|0.00	/	1.00|0.00	1.00|0.00
500	100	1.10|0.00	1.00|0.00	0.00|0.00	1.00|0.00	1.00|0.00	1.10|4.00	2.81|0.00	1.18|0.00	1.04|0.00	1.00|0.00	1.00|0.00	1.00|0.00
	200	1.00|0.00	1.00|0.00	0.00|0.00	1.00|0.00	1.00|0.00	1.09|4.00	3.00|0.00	2.80|0.00	1.00|0.00	1.00|0.00	1.00|0.00	0.00|0.00
	500	1.00|0.00	1.00|0.00	0.00|0.00	1.00|0.00	1.00|0.00	1.15|4.00	3.00|0.00	3.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00	0.00|0.00
	2000	1.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.29|0.00	3.00|0.00	3.00|0.00	1.05|0.00	/	1.00|0.00	1.00|0.00
2000	100	2.89|0.00	0.00|0.00	0.00|0.00	0.00|0.00	0.00|0.00	1.00|4.00	3.00|0.00	2.35|0.00	1.03|0.00	1.00|0.00	0.00|0.00	0.00|0.00
	200	2.48|0.00	2.00|0.00	1.00|0.00	2.00|0.00	2.00|0.00	1.00|4.00	3.00|0.00	3.00|0.00	1.04|0.00	1.00|0.00	1.00|0.00	0.00|0.00
	500	1.17|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00	3.00|0.00	3.00|0.00	1.09|0.00	1.00|0.00	1.00|0.00	1.00|0.00
	2000	1.06|0.00	1.06|0.00	1.00|0.00	1.06|0.00	1.10|0.00	1.00|4.00	3.00|0.00	3.00|0.00	1.20|0.00	/	1.00|0.00	1.14|0.00

NOTE: The upper panel shows the frequencies of correct selection, over-selection, and under-selection, while the lower panel reports the mean of the numbers of over-selected and under-selected factors.

mean squared error. (b) CV_2: Wold-style cross-validation (Wold 1978), which randomly partitions entries of the data matrix into folds. It treats the entries in the held-out data as missing and imputes them using a singular value decomposition-based method (Troyanskaya et al. 2001). It then fits the model to the imputed data and calculates the mean squared error on the held-out data. (c) CV_3: modified Wold-style cross-validation, which is similar to CV_2 but uses the algorithm proposed by Chen and Li (2022) for training on the held-in data and then evaluate the performance in terms of the likelihood function on the held-out data. For CV_1, we partitioned the data matrix into four submatrices (i.e., 4 folds) by dividing the rows and columns into two folds each, and for both CV_2 and CV_3, we adopted the standard 5-fold partitioning. Also included are a ratio-based selection rule (RR) from Jentsch, Lee, and Mammen (2021) and the Gibbs sampling-based adaptive method for selecting the number of factors (adaGibbs) proposed by Bhattacharya and Dunson (2011). Implementation details of these criteria can be found in the supplementary material.

Example 1. Data were generated from the Poisson factor model (1). The true number of factors was set to 5. The factors f_{ik} were generated independently from the standard normal distribution, $N(0, 1)$, and the intercepts α_j and loadings l_{jk} were independently drawn uniformly on the interval $(-1, 1)$.

Example 2. The same as Example 1, except that the fifth factor f_{i5} was generated from $N(0, 0.5)$, making its detection more challenging than the first four factors.

Example 3. The same as Example 1, except that 30% of the counts were generated from a negative binomial distribution. The dispersion parameter was randomly selected from $\{0.1, 1, 10\}$. This example examines the robustness of each criterion against model mis-specification.

Example 4. The same as Example 1, except that 10% of the features underwent contamination by doubling the natural parameter of the corresponding Poisson distribution. This introduces difficulty in distinguishing between true signals and noise.

Example 5. The same as Example 1, except that the observed values x_{ij} are subject to random missingness with a missing probability of 50%. This explores the potential effect of missing data. Currently, the estimation procedure we employ cannot handle missing observations. Directly addressing this issue would require extending both the estimation procedure and the proposed criteria, as well as establishing their theoretical results in the presence of missing data. While this represents an important and interesting direction of future work, it is beyond the scope of the current study. We simply imputed missing values with zeros, as suggested by Jin, Miao, and Su (2021).

Table 2. Performance of factor selection based on 100 data repetitions for various criteria in **Example 2**.

		IC_1	IC_2	IC_3	IC_4	IC_5	RR	CV_1	CV_2	CV_3	adaGibbs	VSDT	ICDT
n	*p*	Frequency of correct (over|under) selection											
100	100	96 (3|1)	96 (3|1)	14 (0|86)	96 (3|1)	97 (2|1)	24 (23|53)	0 (100|0)	0 (100|0)	58 (42|0)	63 (36|1)	97 (2|1)	98 (0|2)
	200	95 (5|0)	87 (13|0)	66 (0|34)	95 (5|0)	96 (4|0)	15 (30|55)	0 (100|0)	0 (100|0)	60 (40|0)	59 (41|0)	96 (4|0)	99 (0|1)
	500	95 (5|0)	86 (14|0)	87 (0|13)	95 (5|0)	95 (5|0)	14 (34|52)	0 (100|0)	0 (100|0)	49 (51|0)	52 (48|0)	94 (6|0)	100 (0|0)
	2000	92 (8|0)	0 (100|0)	90 (0|10)	92 (8|0)	94 (6|0)	4 (40|56)	0 (100|0)	0 (100|0)	54 (46|0)	/	94 (6|0)	97 (3|0)
200	100	96 (4|0)	100 (0|0)	76 (0|24)	100 (0|0)	100 (0|0)	54 (25|21)	0 (100|0)	0 (100|0)	56 (44|0)	51 (49|0)	98 (2|0)	100 (0|0)
	200	96 (4|0)	96 (4|0)	100 (0|0)	96 (4|0)	97 (3|0)	53 (27|20)	0 (100|0)	0 (100|0)	56 (44|0)	58 (42|0)	96 (4|0)	99 (1|0)
	500	97 (3|0)	95 (5|0)	98 (2|0)	97 (3|0)	97 (3|0)	26 (45|29)	0 (100|0)	0 (100|0)	39 (61|0)	58 (42|0)	96 (4|0)	98 (2|0)
	2000	95 (5|0)	26 (74|0)	97 (3|0)	95 (5|0)	95 (5|0)	9 (60|31)	0 (100|0)	0 (100|0)	46 (54|0)	/	94 (6|0)	97 (3|0)
500	100	96 (4|0)	100 (0|0)	96 (0|4)	100 (0|0)	100 (0|0)	70 (25|5)	0 (100|0)	0 (100|0)	51 (49|0)	56 (44|0)	100 (0|0)	100 (0|0)
	200	97 (3|0)	99 (1|0)	100 (0|0)	99 (1|0)	100 (0|0)	84 (8|8)	0 (100|0)	0 (100|0)	42 (58|0)	57 (43|0)	100 (0|0)	100 (0|0)
	500	95 (5|0)	95 (5|0)	98 (2|0)	95 (5|0)	96 (4|0)	64 (28|8)	0 (100|0)	0 (100|0)	41 (59|0)	53 (47|0)	95 (5|0)	98 (2|0)
	2000	87 (13|0)	61 (39|0)	98 (2|0)	87 (13|0)	97 (3|0)	60 (31|9)	0 (100|0)	0 (100|0)	24 (76|0)	/	94 (6|0)	97 (3|0)
2000	100	0 (100|0)	99 (0|1)	99 (0|1)	99 (0|1)	99 (0|1)	90 (10|0)	0 (100|0)	0 (100|0)	30 (70|0)	64 (36|0)	100 (0|0)	99 (0|1)
	200	2 (98|0)	96 (4|0)	97 (3|0)	96 (4|0)	96 (4|0)	96 (2|2)	0 (100|0)	0 (100|0)	18 (82|0)	51 (49|0)	97 (3|0)	98 (2|0)
	500	92 (8|0)	96 (4|0)	97 (3|0)	96 (4|0)	96 (4|0)	96 (3|1)	0 (100|0)	0 (100|0)	15 (85|0)	54 (46|0)	98 (2|0)	99 (1|0)
	2000	90 (10|0)	90 (10|0)	93 (7|0)	90 (10|0)	92 (8|0)	91 (9|0)	0 (100|0)	0 (100|0)	16 (84|0)	/	87 (13|0)	93 (7|0)
n	*p*	Mean of the number of over|under selected factors											
100	100	1.00|1.00	1.00|1.00	0.00|1.06	1.00|1.00	1.00|1.00	1.22|3.72	1.11|0.00	1.01|0.00	1.14|0.00	1.00|1.00	1.00|1.00	0.00|1.00
	200	1.00|0.00	1.00|0.00	0.00|1.00	1.00|0.00	1.00|0.00	1.37|3.80	1.40|0.00	1.01|0.00	1.10|0.00	1.00|0.00	1.00|0.00	0.00|1.00
	500	1.00|0.00	1.14|0.00	0.00|1.00	1.00|0.00	1.00|0.00	1.53|3.60	2.57|0.00	1.12|0.00	1.12|0.00	1.00|0.00	1.00|0.00	0.00|0.00
	2000	1.00|0.00	2.80|0.00	0.00|1.00	1.00|0.00	1.00|0.00	1.65|3.71	3.00|0.00	1.90|0.00	1.15|0.00	/	1.00|0.00	1.00|0.00
200	100	1.00|0.00	0.00|0.00	0.00|1.00	0.00|0.00	0.00|0.00	1.08|3.95	1.38|0.00	1.00|0.00	1.14|0.00	1.00|0.00	1.00|0.00	0.00|0.00
	200	1.00|0.00	1.00|0.00	0.00|0.00	1.00|0.00	1.00|0.00	1.30|3.90	2.38|0.00	1.27|0.00	1.20|0.00	1.00|0.00	1.00|0.00	1.00|0.00
	500	1.00|0.00	1.20|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.31|3.97	3.00|0.00	2.62|0.00	1.07|0.00	1.00|0.00	1.00|0.00	1.00|0.00
	2000	1.00|0.00	1.82|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.50|3.84	3.00|0.00	3.00|0.00	1.11|0.00	/	1.00|0.00	1.00|0.00
500	100	1.00|0.00	0.00|0.00	0.00|1.00	0.00|0.00	0.00|0.00	1.16|4.00	2.60|0.00	1.13|0.00	1.04|0.00	1.00|0.00	0.00|0.00	0.00|0.00
	200	1.00|0.00	1.00|0.00	0.00|0.00	1.00|0.00	0.00|0.00	1.00|4.00	3.00|0.00	2.67|0.00	1.05|0.00	1.00|0.00	0.00|0.00	0.00|0.00
	500	1.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.11|4.00	3.00|0.00	3.00|0.00	1.02|0.00	1.00|0.00	1.00|0.00	1.00|0.00
	2000	1.15|0.00	1.38|0.00	1.00|0.00	1.15|0.00	1.33|0.00	1.13|4.00	3.00|0.00	3.00|0.00	1.18|0.00	/	1.33|0.00	1.33|0.00
2000	100	2.96|0.00	0.00|1.00	0.00|1.00	0.00|1.00	0.00|1.00	1.00|0.00	3.00|0.00	1.98|0.00	1.01|0.00	1.00|0.00	0.00|0.00	0.00|1.00
	200	2.70|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.00|4.00	3.00|0.00	3.00|0.00	1.10|0.00	1.00|0.00	1.00|0.00	1.00|0.00
	500	1.12|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.00|4.00	3.00|0.00	3.00|0.00	1.08|0.00	1.00|0.00	1.00|0.00	1.00|0.00
	2000	1.10|0.00	1.10|0.00	1.14|0.00	1.10|0.00	1.12|0.00	1.00|0.00	3.00|0.00	3.00|0.00	1.07|0.00	/	1.08|0.00	1.14|0.00

NOTE: The upper panel shows the frequencies of correct selection, over-selection, and under-selection, while the lower panel reports the mean of the numbers of over-selected and under-selected factors.

For each example, we set $n, p \in \{100, 200, 500, 2000\}$, resulting in a total of 16 data configurations. For each configuration, we generated 100 simulated datasets. We evaluated the performance of each criterion by calculating the frequencies of correct selection, under-selection, and over-selection of the true number of factors as well as the mean of the numbers of over-selected and under-selected factors.

The simulation results are shown in Tables 1–5. From Tables 1 and 2, it is evident that when the model was correctly specified, ICDT consistently had the best performance. The VSDT criterion tended to slightly over-select the number of factors. There was a clear tendency of over-selection for IC_1 and IC_2 for certain configurations of n and p. In contrast, the behavior of IC_3 was different, with a tendency to under-select the factors in the presence of a weak factor. On the other hand, IC_4 and IC_5 performed slightly worse than ICDT. From Tables 3 and 4, it is apparent that the performance of all criteria deteriorated in the presence of model mis-specification or data contamination. Nevertheless, ICDT displayed a certain level of robustness and demonstrated superior or competitive performance compared to other criteria. IC_1, IC_2, and the VSDT criterion all had a strong tendency toward over-selection. This trend was also observed for IC_4 when the model was incorrectly specified. From Table 5, we see that while the presence of missing data deteriorates the performance of all methods, ICDT demonstrates a clear advantage in controlling factor overestimation compared to traditional information criteria and VSDT.

The simulation results show that CV_3 and adaGibbs exhibited performance characteristics similar to VSDT, though slightly inferior. Meanwhile, CV_1 and CV_2 overestimated the number of factors in most settings, and the ratio-based rule performed poorly, with unstable performance across different configurations of n and p.

We also examined additional simulated examples: (a) the true number of factors was set to 2, (b) the factors were sampled independently from a uniform distribution over $(-2, 2)$, and (c) 10% of the samples underwent contamination by doubling the natural parameter of the corresponding Poisson distribution. The observations from these additional examples were qualitatively the same. See the supplementary material for detailed results.

In the supplementary material, we further assessed the performance of various methods in estimating factors and loadings. Simulation results show that under-selection results in significantly greater estimation error than over-selection. Additionally, to explore the impact of the number of data repetitions, we increased it from 100 to 1000. The conclusions remain consistent with those obtained using only 100 repetitions.

So far, we adopted $f(n, p) = \ln\{np/(n + p)\}$, as proposed by Bai and Ng (2002), which has demonstrated good performance in empirical applications (Hallin and Liška 2007; Jentsch, Lee,

Table 3. Performance of factor selection based on 100 data repetitions for various criteria in **Example 3**.

		IC_1	IC_2	IC_3	IC_4	IC_5	RR	CV_1	CV_2	CV_3	adaGibbs	VSDT	ICDT
n	p	Frequency of correct (over|under) selection											
100	100	9 (91|0)	9 (91|0)	78 (6|16)	9 (91|0)	73 (27|0)	11 (13|76)	0 (100|0)	0 (100|0)	41 (59|0)	66 (34|0)	46 (54|0)	90 (10|0)
	200	25 (75|0)	0 (100|0)	92 (7|1)	25 (75|0)	64 (36|0)	10 (15|75)	0 (100|0)	0 (100|0)	52 (48|0)	62 (38|0)	22 (78|0)	89 (11|0)
	500	57 (43|0)	0 (100|0)	88 (10|2)	57 (43|0)	68 (32|0)	12 (21|67)	0 (100|0)	0 (100|0)	44 (56|0)	57 (43|0)	43 (57|0)	74 (26|0)
	2000	63 (37|0)	0 (100|0)	69 (27|4)	63 (37|0)	64 (36|0)	17 (18|65)	0 (100|0)	0 (100|0)	54 (46|0)	/	26 (73|1)	74 (25|1)
200	100	0 (100|0)	14 (86|0)	87 (11|2)	14 (86|0)	65 (35|0)	35 (10|55)	0 (100|0)	0 (100|0)	50 (50|0)	65 (35|0)	13 (87|0)	82 (18|0)
	200	0 (100|0)	0 (100|0)	86 (14|0)	0 (100|0)	55 (45|0)	29 (20|51)	0 (100|0)	0 (100|0)	58 (42|0)	52 (48|0)	0 (100|0)	83 (17|0)
	500	31 (69|0)	0 (100|0)	84 (16|0)	31 (69|0)	63 (37|0)	22 (25|53)	0 (100|0)	0 (100|0)	54 (46|0)	65 (35|0)	1 (99|0)	80 (20|0)
	2000	11 (89|0)	0 (100|0)	81 (19|0)	11 (89|0)	53 (47|0)	8 (31|61)	0 (100|0)	0 (100|0)	60 (40|0)	/	0 (100|0)	76 (24|0)
500	100	0 (100|0)	53 (47|0)	96 (4|0)	53 (47|0)	77 (23|0)	62 (16|22)	0 (100|0)	0 (100|0)	69 (31|0)	53 (47|0)	2 (98|0)	94 (6|0)
	200	0 (100|0)	22 (78|0)	95 (5|0)	22 (78|0)	78 (22|0)	48 (21|31)	0 (100|0)	0 (100|0)	75 (25|0)	58 (42|0)	1 (99|0)	93 (7|0)
	500	8 (92|0)	8 (92|0)	93 (7|0)	8 (92|0)	85 (15|0)	54 (20|26)	0 (100|0)	0 (100|0)	67 (33|0)	61 (39|0)	2 (98|0)	92 (8|0)
	2000	67 (33|0)	2 (98|0)	94 (6|0)	67 (33|0)	71 (29|0)	70 (11|19)	0 (100|0)	0 (100|0)	55 (45|0)	/	6 (94|0)	93 (7|0)
2000	100	0 (100|0)	93 (7|0)	97 (3|0)	93 (7|0)	93 (7|0)	89 (6|5)	0 (100|0)	0 (100|0)	83 (17|0)	59 (41|0)	0 (100|0)	94 (6|0)
	200	0 (100|0)	92 (8|0)	98 (2|0)	92 (8|0)	92 (8|0)	87 (9|4)	0 (100|0)	0 (100|0)	66 (34|0)	65 (35|0)	10 (89|1)	92 (8|0)
	500	2 (98|0)	85 (15|0)	88 (12|0)	85 (15|0)	85 (15|0)	79 (14|7)	0 (100|0)	0 (100|0)	38 (62|0)	58 (42|0)	18 (82|0)	86 (14|0)
	2000	66 (34|0)	66 (34|0)	84 (16|0)	66 (34|0)	72 (28|0)	89 (7|4)	0 (100|0)	0 (100|0)	17 (83|0)	/	31 (69|0)	75 (25|0)
n	p	Mean of the number of over|under selected factors											
100	100	1.91|0.00	1.91|0.00	1.00|1.00	1.91|0.00	1.15|0.00	1.54|3.68	1.06|0.00	1.00|0.00	1.63|0.00	1.00|0.00	1.09|0.00	1.00|0.00
	200	1.47|0.00	2.73|0.00	1.00|1.00	1.47|0.00	1.17|0.00	1.07|3.65	1.02|0.00	1.00|0.00	1.58|0.00	1.00|0.00	1.38|0.00	1.09|0.00
	500	1.33|0.00	2.87|0.00	1.20|1.00	1.33|0.00	1.34|0.00	1.19|3.75	1.35|0.00	1.00|0.00	1.50|0.00	1.00|0.00	1.67|0.00	1.27|0.00
	2000	1.30|0.00	2.85|0.00	1.07|1.00	1.30|0.00	1.31|0.00	1.33|3.66	2.99|0.00	1.00|0.00	1.26|0.00	/	1.53|3.00	1.72|1.00
200	100	2.84|0.00	1.53|0.00	1.00|1.00	1.53|0.00	1.14|0.00	1.00|3.82	1.03|0.00	1.00|0.00	1.54|0.00	1.00|0.00	1.82|0.00	1.06|0.00
	200	2.60|0.00	2.60|0.00	1.07|0.00	2.60|0.00	1.27|0.00	1.10|3.69	1.04|0.00	1.00|0.00	1.38|0.00	1.00|0.00	2.14|0.00	1.00|0.00
	500	1.54|0.00	2.80|0.00	1.06|0.00	1.54|0.00	1.22|0.00	1.20|3.70	2.21|0.00	1.01|0.00	1.37|0.00	1.00|0.00	2.20|0.00	1.15|0.00
	2000	1.43|0.00	2.80|0.00	1.05|0.00	1.43|0.00	1.17|0.00	1.16|3.70	3.00|0.00	2.14|0.00	1.25|0.00	/	2.19|0.00	1.17|0.00
500	100	2.93|0.00	1.19|0.00	1.00|0.00	1.19|0.00	1.09|0.00	1.00|3.91	1.21|0.00	1.00|0.00	1.19|0.00	1.00|0.00	2.73|0.00	1.17|0.00
	200	2.91|0.00	1.97|0.00	1.20|0.00	1.97|0.00	1.09|0.00	1.14|3.84	1.94|0.00	1.01|0.00	1.04|0.00	1.00|0.00	2.86|0.00	1.14|0.00
	500	2.47|0.00	2.47|0.00	1.00|0.00	2.47|0.00	1.00|0.00	1.10|3.88	3.00|0.00	2.58|0.00	1.09|0.00	1.00|0.00	2.58|0.00	1.12|0.00
	2000	1.48|0.00	2.78|0.00	2.00|0.00	1.48|0.00	1.48|0.00	1.00|4.00	3.00|0.00	3.00|0.00	1.16|0.00	/	2.13|0.00	1.57|0.00
2000	100	2.93|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.00|4.00	2.91|0.00	1.00|0.00	1.06|0.00	1.00|0.00	2.84|0.00	1.00|0.00
	200	2.94|0.00	1.12|0.00	1.00|0.00	1.25|0.00	1.12|0.00	1.00|4.00	3.00|0.00	1.96|0.00	1.00|0.00	1.00|0.00	2.90|2.00	1.00|0.00
	500	2.70|0.00	1.33|0.00	1.33|0.00	1.33|0.00	1.27|0.00	1.14|4.00	3.00|0.00	3.00|0.00	1.02|0.00	1.00|0.00	2.59|0.00	1.29|0.00
	2000	1.26|0.00	1.26|0.00	1.69|0.00	1.26|0.00	1.29|0.00	1.00|4.00	3.00|0.00	3.00|0.00	1.00|0.00	/	1.62|0.00	1.32|0.00

NOTE: The upper panel shows the frequencies of correct selection, over-selection, and under-selection, while the lower panel reports the mean of the numbers of over-selected and under-selected factors.

Table 4. Performance of factor selection based on 100 data repetitions for various criteria in **Example 4**.

		IC_1	IC_2	IC_3	IC_4	IC_5	RR	CV_1	CV_2	CV_3	adaGibbs	VSDT	ICDT
n	*p*	Frequency of correct (over|under) selection											
100	100	94 (6|0)	94 (6|0)	84 (0|16)	94 (6|0)	94 (6|0)	40 (19|41)	0 (100|0)	0 (100|0)	75 (25|0)	56 (44|0)	96 (4|0)	98 (1|1)
	200	96 (4|0)	94 (6|0)	97 (2|1)	96 (4|0)	97 (3|0)	23 (24|53)	0 (100|0)	0 (100|0)	81 (19|0)	56 (44|0)	97 (3|0)	98 (2|0)
	500	94 (6|0)	83 (17|0)	96 (4|0)	94 (6|0)	94 (6|0)	9 (42|49)	0 (100|0)	0 (100|0)	82 (18|0)	65 (35|0)	91 (9|0)	96 (4|0)
	2000	87 (13|0)	0 (100|0)	96 (4|0)	87 (13|0)	89 (11|0)	2 (30|68)	0 (100|0)	0 (100|0)	72 (28|0)	/	88 (12|0)	93 (7|0)
200	100	93 (7|0)	95 (5|0)	98 (1|1)	95 (5|0)	98 (2|0)	57 (27|16)	0 (100|0)	0 (100|0)	79 (21|0)	58 (42|0)	98 (2|0)	99 (1|0)
	200	99 (1|0)	99 (1|0)	100 (0|0)	99 (1|0)	99 (1|0)	56 (29|15)	0 (100|0)	0 (100|0)	85 (15|0)	52 (48|0)	97 (3|0)	99 (1|0)
	500	81 (19|0)	70 (30|0)	94 (6|0)	81 (19|0)	83 (17|0)	27 (49|24)	0 (100|0)	0 (100|0)	82 (18|0)	56 (44|0)	78 (22|0)	92 (8|0)
	2000	80 (20|0)	16 (84|0)	93 (7|0)	80 (20|0)	81 (19|0)	18 (54|28)	0 (100|0)	0 (100|0)	72 (28|0)	/	76 (24|0)	91 (9|0)
500	100	92 (8|0)	98 (2|0)	99 (1|0)	98 (2|0)	98 (2|0)	82 (12|6)	0 (100|0)	0 (100|0)	88 (12|0)	56 (44|0)	88 (12|0)	98 (2|0)
	200	75 (25|0)	79 (21|0)	97 (3|0)	79 (21|0)	89 (11|0)	88 (10|2)	0 (100|0)	0 (100|0)	80 (20|0)	57 (43|0)	78 (22|0)	90 (10|0)
	500	85 (15|0)	85 (15|0)	97 (3|0)	85 (15|0)	87 (13|0)	65 (30|5)	0 (100|0)	0 (100|0)	77 (23|0)	53 (47|0)	63 (37|0)	88 (12|0)
	2000	64 (36|0)	47 (53|0)	86 (14|0)	64 (36|0)	64 (36|0)	58 (34|8)	0 (100|0)	0 (100|0)	65 (35|0)	/	56 (44|0)	75 (25|0)
2000	100	0 (100|0)	100 (0|0)	100 (0|0)	100 (0|0)	100 (0|0)	91 (7|2)	0 (100|0)	0 (100|0)	85 (15|0)	62 (38|0)	99 (1|0)	100 (0|0)
	200	1 (99|0)	82 (18|0)	95 (5|0)	82 (18|0)	82 (18|0)	75 (23|2)	0 (100|0)	0 (100|0)	65 (35|0)	62 (38|0)	76 (24|0)	89 (11|0)
	500	9 (91|0)	62 (38|0)	85 (15|0)	62 (38|0)	67 (33|0)	66 (34|0)	0 (100|0)	0 (100|0)	55 (45|0)	58 (42|0)	61 (39|0)	81 (19|0)
	2000	16 (84|0)	16 (84|0)	59 (41|0)	16 (84|0)	43 (57|0)	73 (26|1)	0 (100|0)	0 (100|0)	28 (72|0)	/	32 (68|0)	71 (29|0)
n	*p*	Mean of the number of over|under selected factors											
100	100	1.00|0.00	1.00|0.00	0.00|1.00	1.00|0.00	1.00|0.00	1.11|3.76	1.15|0.00	1.00|0.00	1.16|0.00	1.00|0.00	1.00|0.00	1.00|1.00
	200	1.00|0.00	1.00|0.00	1.00|1.00	1.00|0.00	1.00|0.00	1.42|3.72	1.53|0.00	1.01|0.00	1.05|0.00	1.00|0.00	1.33|0.00	1.00|0.00
	500	1.00|0.00	1.12|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.36|3.80	2.78|0.00	1.23|0.00	1.11|0.00	1.00|0.00	1.00|0.00	1.00|0.00
	2000	1.00|0.00	2.73|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.73|3.85	3.00|0.00	2.18|0.00	1.14|0.00	/	1.00|0.00	1.00|0.00
200	100	1.00|0.00	1.00|0.00	1.00|1.00	1.00|0.00	1.00|0.00	1.11|3.94	1.40|0.00	1.00|0.00	1.19|0.00	1.00|0.00	1.00|0.00	1.00|0.00
	200	1.00|0.00	1.00|0.00	0.00|0.00	1.00|0.00	1.00|0.00	1.10|3.87	2.59|0.00	1.28|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00
	500	1.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.27|3.88	3.00|0.00	2.83|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00
	2000	1.00|0.00	1.73|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.46|3.89	3.00|0.00	3.00|0.00	1.18|0.00	/	1.04|0.00	1.11|0.00
500	100	1.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.00|4.00	2.74|0.00	1.26|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00
	200	1.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.00|0.00	1.00|4.00	3.00|0.00	2.74|0.00	1.10|0.00	1.00|0.00	1.00|0.00	1.00|0.00
	500	1.07|0.00	1.07|0.00	1.00|0.00	1.07|0.00	1.00|0.00	1.07|3.80	3.00|0.00	3.00|0.00	1.17|0.00	1.00|0.00	1.00|0.00	1.00|0.00
	2000	1.08|0.00	1.11|0.00	1.00|0.00	1.08|0.00	1.08|0.00	1.12|3.88	3.00|0.00	3.00|0.00	1.11|0.00	/	1.07|0.00	1.04|0.00
2000	100	2.94|0.00	0.00|0.00	0.00|0.00	0.00|0.00	0.00|0.00	1.00|4.00	3.00|0.00	2.31|0.00	1.07|0.00	1.00|0.00	1.00|0.00	0.00|0.00
	200	2.53|0.00	1.06|0.00	1.00|0.00	1.06|0.00	1.06|0.00	1.13|4.00	3.00|0.00	3.00|0.00	1.06|0.00	1.00|0.00	1.04|0.00	1.09|0.00
	500	1.58|0.00	1.11|0.00	1.00|0.00	1.11|0.00	1.06|0.00	1.09|0.00	3.00|0.00	3.00|0.00	1.16|0.00	1.00|0.00	1.18|0.00	1.00|0.00
	2000	1.12|0.00	1.12|0.00	1.17|0.00	1.12|0.00	1.12|0.00	1.00|4.00	3.00|0.00	3.00|0.00	1.36|0.00	/	1.15|0.00	1.10|0.00

NOTE: The upper panel shows the frequencies of correct selection, over-selection, and under-selection, while the lower panel reports the mean of the numbers of over-selected and under-selected factors.

and Mammen 2021; Liu et al. 2023). However, the choice of $f(n,p)$ remains somewhat ad hoc. On one hand, the consistency of the proposed criterion ICDT holds under a wide range of choices for $f(n,p)$, as long as it diverges slowly as n and p both tend to infinity. On the other hand, multiplying $f(n,p)$ by a constant c does not affect the consistency of the resulting criterion. With $c = 1$, the reported results show that ICDT performs comparably well with existing criteria. Furthermore, the upper panel in Figure 1 demonstrates that ICDT remains robust over a broad range of c values, indicating relative insensitivity to its choice. Inspired by Hallin and Liška (2007) and Alessi, Barigozzi, and Capasso (2010), we explored a method for determining c and $f(n,p)$ by treating data thinning as a form of information splitting (Leiner et al. 2023; Neufeld et al. 2024). Unlike sample splitting, which is a discrete process, data thinning is a continuous process indexed by the thinning parameter π, which takes continuous values between 0 and 1. This provides a natural framework for adjusting data information. Specifically, we examined the stability of the ICDT criterion with respect to c and/or $f(n,p)$ by varying the thinning parameter.

Consider a sequence of thinning parameters $\pi^0 = 0 < \pi^1 < \cdots < \pi^H = 1$. For each $h \in \{0,\ldots,H\}$, we apply ICDT and calculate the estimated number of factors $\hat{r}_{\pi^h}$, which is a nonincreasing function of c. Clearly, different values of c lead to different behaviors of $\hat{r}_{\pi^h}$ as a function of h. The stability of ICDT with respect to c can be measured by the empirical variance of $\hat{r}_{\pi^h}$ as a function of h, defined as

$$S_c = \frac{1}{H}\sum_{h=1}^{H}\left(\hat{r}_{\pi^h} - \frac{1}{H}\sum_{h=1}^{H}\hat{r}_{\pi^h}\right)^2.$$

In our implementation, we examined $c \in \{0, 0.1, 0.5, 1, 1.5, 2, 5, 10\}$ and varied π^h from 0.3 to 1 in increments of 0.05. The middle panel in Figure 1 displays the average estimated number of factors across 100 simulations for different values of c as π varies. We see that $c = 1$ appears to be a good choice, even with relatively limited information. The lower panel in Figure 1 illustrates the stability and accuracy of the ICDT criterion as c varies. We observe that values of c near 1 represent a robust choice, as they provide both accurate estimates and high stability.

6. Real Data Analysis

We illustrate the use of the ICDT criterion by applying it to two real datasets.

6.1. BBC News Data

We applied the proposed criteria to a text corpus previously used in the evaluation of document clustering algorithms (Greene and Cunningham 2006; Greene, O'Callaghan, and Cunningham 2014). The corpus comprised 2225 general BBC news articles

Table 5. Performance of factor selection based on 100 data repetitions for various criteria in **Example 5**. The upper panel shows the frequencies of correct selection, over-selection, and under-selection, while the lower panel reports the mean of the numbers of over-selected and under-selected factors.

		IC_1	IC_2	IC_3	IC_4	IC_5	RR	CV_1	CV_2	CV_3	adaGibbs	VSDT	ICDT
n	*p*	Frequency of correct (over|under) selection											
100	100	0 (100|0)	0 (100|0)	9 (0|91)	0 (100|0)	17 (83|0)	0 (0|100)	39 (23|38)	5 (1|94)	24 (24|52)	4 (1|95)	58 (27|15)	33 (2|65)
	200	1 (99|0)	0 (100|0)	39 (3|58)	1 (99|0)	12 (88|0)	1 (0|99)	19 (81|0)	38 (43|19)	43 (31|26)	11 (1|88)	21 (78|1)	57 (18|25)
	500	3 (97|0)	0 (100|0)	63 (11|26)	3 (97|0)	14 (86|0)	5 (2|93)	0 (100|0)	1 (99|0)	48 (47|5)	17 (0|83)	10 (90|0)	54 (38|8)
	2000	10 (90|0)	1 (99|0)	71 (2|27)	10 (90|0)	11 (89|0)	3 (12|85)	0 (100|0)	0 (100|0)	37 (60|3)	/	10 (90|0)	50 (45|5)
200	100	0 (100|0)	0 (100|0)	52 (2|46)	0 (100|0)	8 (92|0)	0 (12|88)	13 (86|1)	39 (38|23)	34 (48|18)	27 (2|71)	9 (91|0)	76 (15|9)
	200	0 (100|0)	0 (100|0)	86 (9|5)	0 (100|0)	4 (96|0)	1 (17|82)	0 (100|0)	0 (100|0)	35 (62|3)	42 (13|45)	0 (100|0)	68 (32|0)
	500	0 (100|0)	0 (100|0)	72 (28|0)	0 (100|0)	3 (97|0)	1 (10|89)	0 (100|0)	0 (100|0)	36 (64|0)	47 (17|36)	0 (100|0)	59 (41|0)
	2000	2 (98|0)	0 (100|0)	80 (20|0)	2 (98|0)	5 (95|0)	7 (3|90)	0 (100|0)	0 (100|0)	41 (59|0)	/	0 (98|2)	62 (38|0)
500	100	0 (100|0)	2 (98|0)	86 (0|14)	2 (98|0)	13 (87|0)	4 (34|62)	0 (100|0)	3 (97|0)	28 (68|4)	41 (23|36)	1 (98|1)	82 (18|0)
	200	0 (100|0)	0 (100|0)	96 (4|0)	0 (100|0)	1 (99|0)	5 (49|46)	0 (100|0)	0 (100|0)	20 (80|0)	48 (41|11)	0 (100|0)	76 (24|0)
	500	0 (100|0)	0 (100|0)	91 (9|0)	0 (100|0)	0 (100|0)	3 (45|52)	0 (100|0)	0 (100|0)	27 (73|0)	54 (45|1)	0 (99|1)	63 (37|0)
	2000	1 (97|2)	1 (97|2)	84 (14|2)	1 (97|2)	1 (97|2)	9 (35|56)	0 (100|0)	0 (100|0)	28 (72|0)	/	3 (88|9)	42 (56|2)
2000	100	0 (100|0)	12 (88|0)	95 (0|5)	12 (88|0)	19 (81|0)	10 (68|22)	0 (100|0)	0 (100|0)	28 (72|0)	55 (28|17)	0 (99|1)	88 (12|0)
	200	0 (100|0)	0 (100|0)	98 (2|0)	0 (100|0)	0 (100|0)	53 (42|5)	0 (100|0)	0 (100|0)	12 (88|0)	64 (34|2)	1 (95|4)	41 (59|0)
	500	1 (99|0)	1 (99|0)	94 (6|0)	1 (99|0)	1 (99|0)	78 (19|3)	0 (100|0)	0 (100|0)	6 (94|0)	57 (43|0)	2 (94|4)	25 (75|0)
	2000	6 (89|5)	6 (89|5)	73 (22|5)	6 (89|5)	6 (89|5)	83 (17|0)	0 (100|0)	0 (100|0)	7 (93|0)	/	8 (48|44)	16 (79|5)
n	*p*	Mean of the number of over|under selected factors											
100	100	2.62|0.00	2.62|0.00	0.00|1.63	2.62|0.00	1.35|0.00	0.00|3.65	1.09|1.29	1.00|2.53	1.62|1.77	1.00|2.16	1.04|1.00	2.00|1.35
	200	2.51|0.00	2.75|0.00	1.00|1.17	2.51|0.00	1.52|0.00	0.00|3.36	1.32|0.00	1.00|1.16	1.19|1.15	1.00|1.78	1.19|1.00	1.44|1.12
	500	1.81|0.00	2.75|0.00	1.00|1.19	1.81|0.00	1.43|0.00	1.50|3.19	1.21|0.00	1.00|0.00	1.15|1.20	0.00|1.52	1.44|0.00	1.47|1.25
	2000	1.64|0.00	2.73|0.00	1.00|1.07	1.64|0.00	1.58|0.00	1.08|2.80	1.76|0.00	1.00|0.00	1.08|1.00	/	1.94|0.00	2.38|1.00
200	100	2.87|0.00	2.64|0.00	1.00|1.13	2.64|0.00	1.51|0.00	1.50|3.72	1.40|1.00	1.00|1.17	1.54|1.06	1.00|1.62	1.54|0.00	1.00|1.00
	200	2.79|0.00	2.79|0.00	1.00|1.00	2.79|0.00	1.90|0.00	1.35|3.59	1.19|0.00	1.00|0.00	1.29|1.00	1.00|1.20	2.46|0.00	1.06|0.00
	500	2.72|0.00	2.91|0.00	1.07|0.00	2.72|0.00	2.01|0.00	1.60|3.54	1.01|0.00	1.00|0.00	1.16|0.00	1.00|1.06	2.17|0.00	1.12|0.00
	2000	2.12|0.00	2.91|0.00	1.00|0.00	2.12|0.00	1.76|0.00	1.67|3.47	1.10|0.00	1.00|0.00	1.10|0.00	/	2.31|1.00	1.66|0.00
500	100	2.93|0.00	2.16|0.00	0.00|1.00	2.16|0.00	1.61|0.00	1.50|3.89	1.62|0.00	1.00|0.00	1.10|1.00	1.00|1.17	2.48|1.00	1.06|0.00
	200	2.94|0.00	2.81|0.00	1.25|0.00	2.81|0.00	2.36|0.00	1.33|3.89	1.00|0.00	1.00|0.00	1.21|0.00	1.00|1.18	2.95|0.00	1.21|0.00
	500	2.88|0.00	2.88|0.00	1.33|0.00	2.88|0.00	2.61|0.00	1.20|3.81	1.00|0.00	1.00|0.00	1.07|0.00	1.00|1.00	2.95|2.00	1.35|0.00
	2000	2.79|2.50	2.89|2.50	1.36|2.50	2.79|2.50	2.60|2.50	1.60|3.82	1.00|0.00	1.00|0.00	1.04|0.00	/	2.93|2.44	1.50|2.50
2000	100	2.98|0.00	1.52|0.00	0.00|1.00	1.52|0.00	1.42|0.00	1.51|4.00	2.08|0.00	1.00|0.00	1.17|0.00	1.00|1.06	2.61|3.00	1.17|0.00
	200	2.89|0.00	2.57|0.00	1.50|0.00	2.57|0.00	2.22|0.00	1.14|4.00	1.01|0.00	1.00|0.00	1.07|0.00	1.00|1.00	2.99|2.00	1.54|0.00
	500	2.92|0.00	2.85|0.00	1.33|0.00	2.85|0.00	2.82|0.00	1.05|4.00	1.00|0.00	1.00|0.00	1.05|0.00	1.00|0.00	2.95|2.25	1.55|0.00
	2000	2.76|2.00	2.76|2.00	1.73|2.00	2.76|2.00	2.65|2.00	1.06|0.00	2.15|0.00	1.00|0.00	1.03|0.00	/	2.79|2.41	1.77|2.00

Table 6. A confusion matrix comparing the assigned classes, which maximize factor scores, to the ground truth document labels for the 2225 articles in the BBC news dataset.

Label	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Business	135	48	0	160	**167**
Entertainment	16	89	**158**	1	122
Politics	**326**	2	75	4	10
Sport	1	0	38	**421**	51
Tech	10	**343**	20	13	15
Top 10 words	govern tax party labour elect plan min-istry Blair UK country	mobile phone music service technology firm company market net user	people go party play BBC lord home work Blair say	game Eng-land player against com-pany club unit wale nation Ireland	sale bank year price film econ-omy Eng-land rate 2004 growth

NOTE: The table also includes the top 10 words with the highest loadings for each factor.

from 2004 to 2005. The raw dataset, along with additional details about the corpus, is available at *http://mlg.ucd.ie/datasets/bbc.html*. We pre-processed the dataset by employing standard stemming and stop-word removal techniques, and excluded terms occurring in fewer than 1% of the documents, resulting in 2913 terms. No further feature selection or term normalization was performed on the term-document matrix.

We set the candidate values for the number of factors in the range $\{1, 2, \ldots, 20\}$, and applied ICDT, VSDT, IC_4, and IC_5 to determine the number of factors. Only ICDT suggested a five-factor structure, which aligns with the fact that each of the 2225 documents corresponds to stories in five topical areas: business, entertainment, politics, sport, and tech (Greene and Cunningham 2006; Greene, O'Callaghan, and Cunningham 2014).

To further check the five-factor structure, we assigned the news articles to five classes maximizing the factor score, and compared the results to the five "ground truth" document labels, resulting in the confusion matrix in Table 6. It is evident that the factors correspond well to the pre-specified labels. Additionally, Table 6 lists the top 10 words with the highest loadings for each factor. The meanings of these words are consistent with the document labels or topical areas.

6.2. Single-Cell RNA Sequencing Data

We further analyzed a dataset of Peripheral Blood Mononuclear Cells (PBMC) freely available from 10X Genomics. This dataset comprises 2700 single cells that were sequenced on the Illumina NextSeq 500 platform. The raw data matrix consists of the numbers of molecules for the genes detected in the cells and can be downloaded from the Seurat tutorial *https://satijalab.org/seurat/articles/pbmc3k_tutorial*.

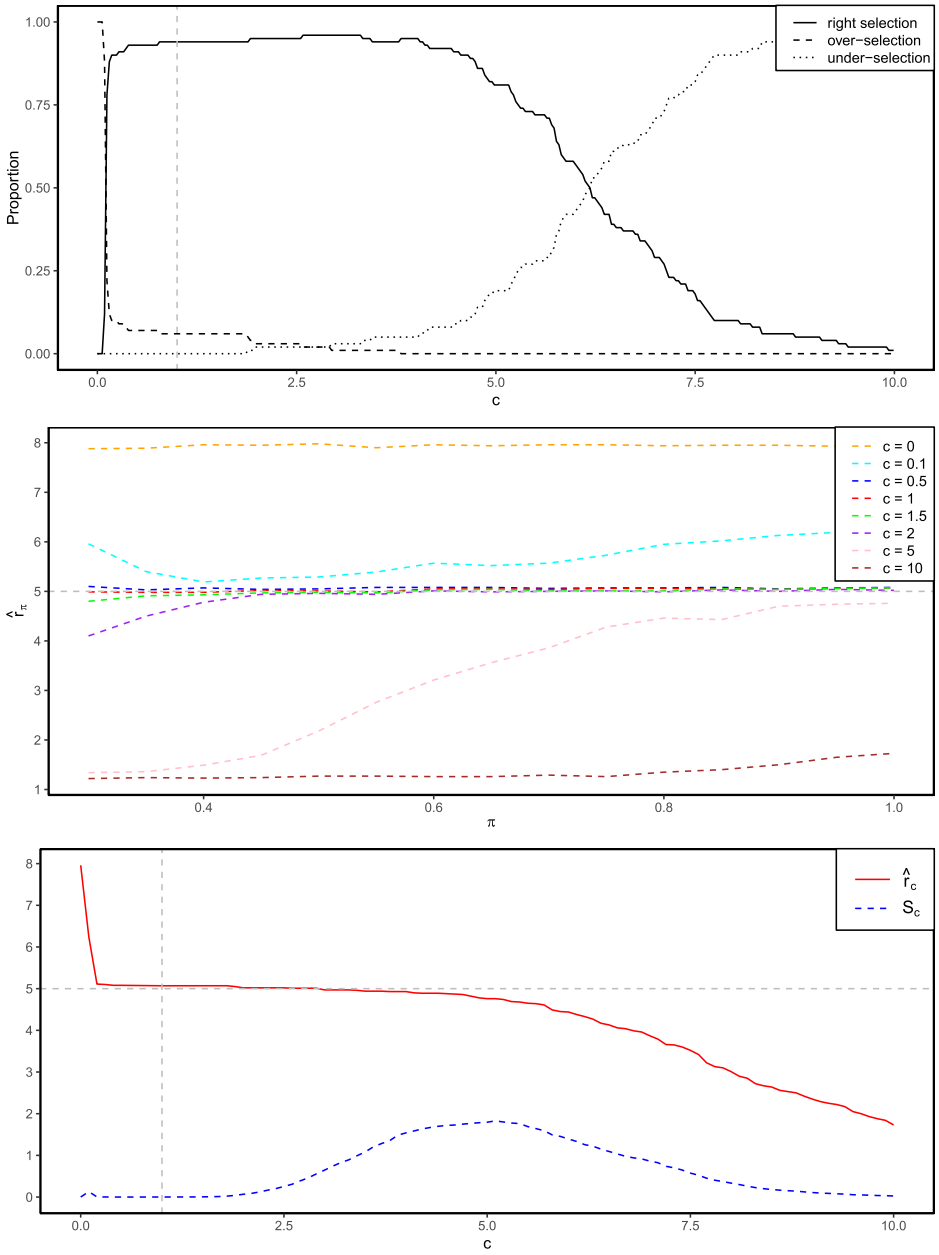

Figure 1. Robustness check of ICDT. The upper panel shows the proportion of correctly estimated, overestimated, and underestimated numbers of factors as a function of c. The middle panel presents the mean of the estimated number of factors as a function of π across different values of c. The lower panel displays the mean estimated number of factors $\hat{r}_c$ and the empirical variance $\hat{S}_c$ as a function of c. The dashed vertical line indicates the value of $c = 1$. The data generation process is the same as Example 1, with $n = 200$, $p = 200$, and the true number of factors set to 5, The results are based on 100 data repetitions.

We then followed the steps described in the tutorial to pre-process the raw count matrix. Specifically, the steps encompass the selection and filtration of cells based on quality control metrics, data normalization and scaling, and the detection of genes that exhibit high cell-to-cell variation (Ilicic et al. 2016; Stuart et al. 2019; Choudhary and Satija 2022). The pre-processed count matrix contains 2638 cells and 2000 genes.

The maximum value of the number of factors was set to 20. The results of applying ICDT, VSDT, IC_4, and IC_5 are presented in Figure 2. We can observe that both ICDT and IC_4 suggest using five factors, which is consistent with previous findings (Neufeld et al. 2024). VSDT appears to be too liberal, retaining as many as 11 factors. The performance of IC_5 is somewhat peculiar because it peaks at 3, contrasting with its relatively liberal behavior compared to ICDT, as observed in the simulations.

For data visualization purposes, we first log-transformed the counts after adding a pseudo-count of one to avoid issues with zeros. For each gene, we then standardized the transformed expression values to have a mean of zero and unit norm. Additionally, for each factor, we ordered both cells and genes according to their factor scores and loadings, respectively. Figure D1 in the supplementary material displays heatmaps for 500 cells (250 most positive and 250 most negative) and 40 genes (20 most positive and 20 most negative). It is evident that the first five factors explain a large proportion of the variability in the data.

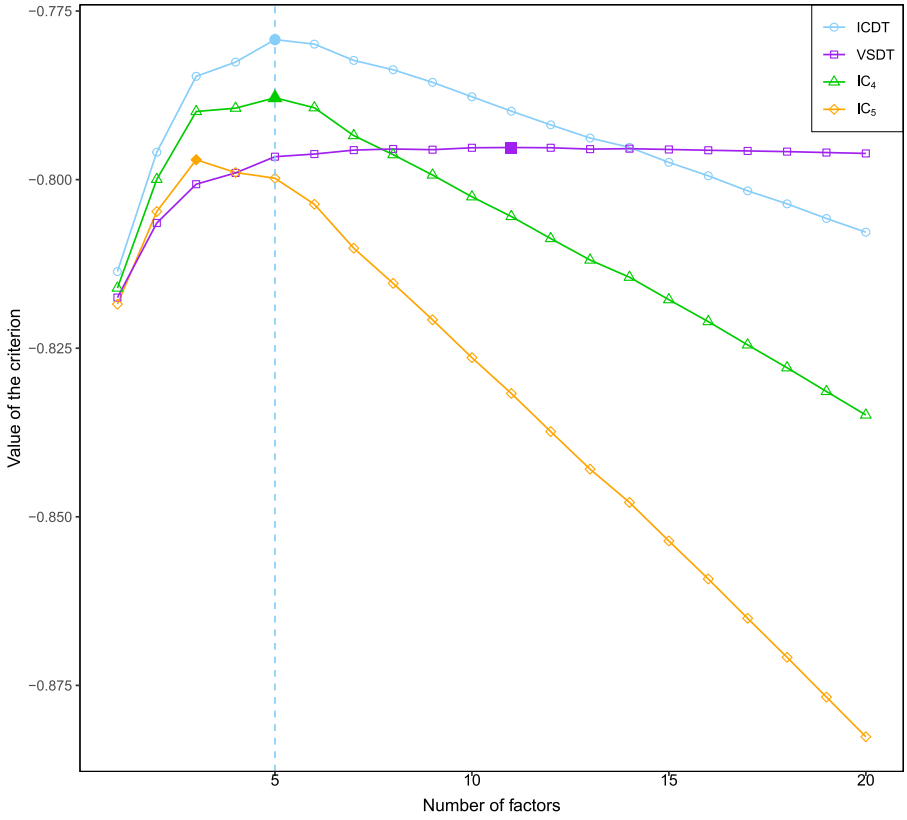

Figure 2. ICDT, VSDT, IC_4, and IC_5 are shown for the best models of each size for the PBMC single-cell RNA sequencing dataset. The criteria have been scaled for display purposes.

7. Discussion

A number of methods have been developed to estimate the factors and loadings in Poisson factor models, or generalized factor models more broadly, with efficient algorithms designed for large-scale applications. However, most existing methods assume the number of factors is known. Although several criteria have been proposed for determining this number, they are often ad hoc or lack theoretical justification. To address this issue, we have introduced two novel and principled criteria, Validation Set via Data Thinning (VSDT) and Information Criterion via Data Thinning (ICDT). These criteria use the thinning property of the Poisson distribution. Unlike data splitting, the advantage of data thinning is that the partition of the count matrix into training and validation sets preserves both the distribution and the underlying data structure.

An important property of VSDT is that the error on the validation dataset can be decomposed into the training error on the training set plus a covariance penalty for the Poisson factor model. This leads directly to a simple and unbiased estimator of the covariance penalty and naturally motivates the development of ICDT. We have derived the theoretical property of the unbiased estimator as well as the selection consistency of ICDT when both the sample size and the number of variables grow to infinity. By incorporating the response inversely into the Poisson factor model, we have introduced the Poisson factor inverse regression model and extended the proposed methods and theory to supervised dimension reduction. We have demonstrated comprehensively the superior performance of ICDT over existing criteria using simulated examples and real datasets.

The proposed methodology still has some limitations that need to be addressed. First, our assumption of independent samples may not always hold true. Relaxing this assumption would broaden the applicability of our criteria, particularly to longitudinal datasets. Second, although our focus has been on large-scale count datasets, our criteria could potentially be applied to other types of datasets, including those with continuous, discrete, or mixed variables. However, this would necessitate theoretical development within a unified framework of generalized factor models, an area we are actively exploring.

The thinning property is not limited to the Poisson distribution and also applies to a range of other distributions (Leiner et al. 2023; Neufeld et al. 2024; Dharamshi et al. 2025). Extending the proposed method to factor models under these distributions presents an interesting direction for future research. For example, it is promising to extend our approach to the negative binomial distribution for modeling count data, which offers greater flexibility than the Poisson distribution and also enjoys the thinning property (Neufeld et al. 2024). Two points are noteworthy. First, both a distributional assumption, such as the Poisson or the negative binomial, and the conditional independence of observations given the latent factors are required. While such assumptions are common in the statistical literature (Chen and Li 2022; Liu et al. 2023), they are less prevalent in the econometric literature (Bai 2003; Jin, Miao, and Su 2021). Second, it would be interesting to explore whether the proposed method can be extended to nonlinear and possibly dynamic factor models (Forni et al. 2000; Wang 2022).

Supplementary Materials

Proofs of theoretical results are given in Section A, implementation details of the algorithm, including pseudo code, are provided in Section B, simulation details along with additional results are presented in Section C, and further results from single-cell RNA sequencing data analysis are included in Section D.

Acknowledgments

We are grateful to the Editor, the Associate Editor, and two anonymous referees for their helpful comments.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

This research was supported in part by the National Natural Science Foundation of China (12222111, 12331009, 12571306), the National Social Science Foundation of China (24BTJ060), the National Key Research and Development Program of China (2024YFA1014101), the Key Laboratory of Scientific and Engineering Computing (Ministry of Education), the Shanghai Frontiers Science Center of Modern Analysis, and the Fundamental Research Funds for the Central Universities.

ORCID

Tao Wang http://orcid.org/0000-0002-1218-4017

References

Ahn, S. C., and Horenstein, A. R. (2013), "Eigenvalue Ratio Test for the Number of Factors," *Econometrica*, 81, 1203–1227. [2]

Alessi, L., Barigozzi, M., and Capasso, M. (2010), "Improved Penalization for Determining the Number of Factors in Approximate Factor Models," *Statistics & Probability Letters*, 80, 1806–1813. [9]

Bai, J. (2003), "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135–171. [12]

Bai, J., and Li, K. (2012), "Statistical Analysis of Factor Models of High Dimension," *The Annals of Statistics*, 40, 436–465. [1,2]

Bai, J., and Ng, S. (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221. [2,4,6,8]

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011), *Latent Variable Models and Factor Analysis: A Unified Approach*, Oxford: Wiley. [1]

Bhattacharya, A., and Dunson, D. B. (2011), "Sparse Bayesian Infinite Factor Models," *Biometrika*, 98, 291–306. [2,7]

Chen, Y., and Li, X. (2022), "Determining the Number of Factors in High-Dimensional Generalized Latent Factor Models," *Biometrika*, 109, 769–782. [1,2,3,4,5,6,7,12]

Choudhary, S., and Satija, R. (2022), "Comparison and Evaluation of Statistical Error Models for scRNA-seq," *Genome Biology*, 23, 1–20. [11]

Cohn, J. B., Liu, Z., and Wardlaw, M. I. (2022), "Count (and Count-Like) Data in Finance," *Journal of Financial Economics*, 146, 529–551. [1]

Cook, R. D., and Forzani, L. (2008), "Principal Fitted Components for Dimension Reduction in Regression," *Statistical Science*, 23, 485–501. [2,5,6]

Dai, C., Lin, B., Xing, X., and Liu, J. S. (2023), "False Discovery Rate Control via Data Splitting," *Journal of the American Statistical Association*, 118, 2503–2520. [3]

Dharamshi, A., Neufeld, A., Motwani, K., Gao, L. L., Witten, D., and Bien, J. (2025), "Generalized Data Thinning Using Sufficient Statistics," *Journal of the American Statistical Association*, 120, 511–523. [3,12]

Efron, B., and Hastie, T. (2016), *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*, Cambridge: Cambridge University Press. [4]

Fan, J., Guo, J., and Zheng, S. (2022), "Estimating Number of Factors by Adjusted Eigenvalues Thresholding," *Journal of the American Statistical Association*, 117, 852–861. [1]

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000), "The Generalized Dynamic-Factor Model: Identification and Estimation," *Review of Economics and Statistics*, 82, 540–554. [12]

Greene, D., and Cunningham, P. (2006), "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering," in *International Conference on Machine Learning*, pp. 377–384, PMLR. [9,10]

Greene, D., O'Callaghan, D., and Cunningham, P. (2014), "How Many Topics? Stability Analysis for Topic Models," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 498–513, Springer. [9,10]

Hallin, M., and Liška, R. (2007), "Determining the Number of Factors in the General Dynamic Factor Model," *Journal of the American Statistical Association*, 102, 603–617. [8,9]

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer. [4]

Ilicic, T., Kim, J. K., Kolodziejczyk, A. A., Bagger, F. O., McCarthy, D. J., Marioni, J. C., and Teichmann, S. A. (2016), "Classification of Low Quality Cells from Single-Cell RNA-Seq Data," *Genome Biology*, 17, 1–15. [11]

Jentsch, C., Lee, E. R., and Mammen, E. (2021), "Poisson Reduced-Rank Models with an Application to Political Text Data," *Biometrika*, 108, 455–468. [2,4,7,9]

Jin, S., Miao, K., and Su, L. (2021), "On Factor Models with Random Missing: EM Estimation, Inference, and Cross Validation," *Journal of Econometrics*, 222, 745–777. [4,7,12]

Kelly, B., Manela, A., and Moreira, A. (2021), "Text Selection," *Journal of Business & Economic Statistics*, 39, 859–879. [1]

Kidziński, Ł., Hui, F. K., Warton, D. I., and Hastie, T. J. (2022), "Generalized Matrix Factorization: Efficient Algorithms for Fitting Generalized Linear Latent Variable Models to Large Data Arrays," *Journal of Machine Learning Research*, 23, 13211–13239. [1,2,3]

Lam, C., Yao, Q., and Bathia, N. (2011), "Estimation of Latent Factors for High-Dimensional Time Series," *Biometrika*, 98, 901–918. [2]

Last, G., and Penrose, M. (2017), *Lectures on the Poisson Process*, Cambridge: Cambridge University Press. [2,3]

Legramanti, S., Durante, D., and Dunson, D. B. (2020), "Bayesian Cumulative Shrinkage for Infinite Factorizations," *Biometrika*, 107, 745–752. [2]

Leiner, J., Duan, B., Wasserman, L., and Ramdas, A. (2023), "Data Fission: Splitting a Single Data Point," *Journal of the American Statistical Association*, 120, 135–146. [2,9,12]

Li, B. (2018), *Sufficient Dimension Reduction: Methods and Applications with R*, Boca Raton, FL: CRC Press. [5]

Li, Q., Cheng, G., Fan, J., and Wang, Y. (2018), "Embracing the Blessing of Dimensionality in Factor Models," *Journal of the American Statistical Association*, 113, 380–389. [1]

Liu, W., Lin, H., Zheng, S., and Liu, J. (2023), "Generalized Factor Model for Ultra-High Dimensional Ccorrelated Variables with Mixed Types," *Journal of the American Statistical Association*, 118, 1385–1401. [1,2,3,4,6,9,12]

Neufeld, A., Dharamshi, A., Gao, L. L., and Witten, D. (2024), "Data Thinning for Convolution-Closed Distributions," *Journal of Machine Learning Research*, 25, 1–35. [2,3,9,11,12]

Owen, A. B., and Wang, J. (2016), "Bi-Cross-Validation for Factor Analysis," *Statistical Science*, 31, 119–139. [3,6]

Pang, D., Zhao, H., and Wang, T. (2024), "Factor Augmented Inverse Regression and its Application to Microbiome Data Analysis," *Journal of the American Statistical Association*, 119, 1957–1967. [2,5]

Rinaldo, A., Wasserman, L., and Gsell, M. (2019), "Bootstrapping and Sample Splitting for High-Dimensional, Assumption-Free Inference," *The Annals of Statistics*, 47, 3438–3469. [3]

Shao, J. (1993), "Linear Model Selection by Cross-Validation," *Journal of the American statistical Association*, 88, 486–494. [4]

Srivastava, S., Engelhardt, B. E., and Dunson, D. B. (2017), "Expandable Factor Analysis," *Biometrika*, 104, 649–663. [2]

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019), "Comprehensive Integration of Single-Cell Data," *Cell*, 177, 1888–1902. [11]

Taddy, M. (2013), "Multinomial Inverse Regression for Text Analysis," *Journal of the American Statistical Association*, 108, 755–770. [2,5]

——— (2015), "Distributed Multinomial Regression," *The Annals of Applied Statistics*, 9, 1394–1414. [2,5]

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001), "Missing Value Estimation Methods for DNA Microarrays," *Bioinformatics*, 17, 520–525. [7]

Wang, F. (2022), "Maximum Likelihood Estimation and Inference for High Dimensional Generalized Factor Models with Application to Factor-Augmented Regressions," *Journal of Econometrics*, 229, 180–200. [1,12]

Wang, J., Huang, M., Torre, E., Dueck, H., Shaffer, S., Murray, J., Raj, A., Li, M., and Zhang, N. R. (2018), "Gene Expression Distribution Deconvolution in Single-Cell RNA Sequencing," *Proceedings of the National Academy of Sciences*, 115, 6437–6446. [1]

Wold, S. (1978), "Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models," *Technometrics*, 20, 397–405. [7]

Zhang, X., Xu, G., and Zhu, J. (2022), "Joint Latent Space Models for Network Data with High-Dimensional Node Variables," *Biometrika*, 109, 707–720. [2]