Psychometrika (2026), **91**, 494–507
doi:10.1017/psy.2026.10092

PSYCHOMETRIC SOCIETY

THEORY AND METHODS

Maximum Softly Penalized Likelihood in Factor Analysis

Philipp Sterzinger[1], Ioannis Kosmidis[2] and Irini Moustaki[1]

[1]Department of Statistics, London School of Economics and Political Science, United Kingdom; [2]Department of Statistics, University of Warwick, United Kingdom

Corresponding author: Philipp Sterzinger; Email: p.sterzinger@lse.ac.uk

(Received 8 October 2025; revised 26 January 2026; accepted 2 February 2026)

Abstract

Estimation in exploratory factor analysis often yields estimates on the boundary of the parameter space. Such occurrences, called Heywood cases, are characterized by non-positive variance estimates and can cause numerical instability, convergence failures, and misleading inferences. We derive sufficient conditions on the model and a penalty to the log-likelihood function that guarantee the existence of maximum penalized likelihood estimates in the interior of the parameter space, and that the corresponding estimators possess desirable asymptotic properties expected by the maximum likelihood estimator, namely, consistency and asymptotic normality. Consistency and asymptotic normality follow when penalization is soft enough, in a way that adapts to the information accumulation about the model parameters. We formally show, for the first time, that the penalties of Akaike (1987, *Psychometrika*, 52, 317–332) and Hirose et al. (2011, *Journal of Data Science*, 9, 243–259) to the log-likelihood of the normal linear factor model satisfy the conditions for existence, and, hence, deal with Heywood cases. Their vanilla versions, though, can result in questionable finite-sample properties in estimation, inference, and model selection. Our maximum softly-penalized likelihood (MSPL) framework ensures that the resulting estimation and inference procedures are asymptotically optimal. Through comprehensive simulation studies and real data analyses, we illustrate the desirable finite-sample properties of the MSPL estimators.

Keywords: Heywood cases; infinite estimates; singular variance components

1. Introduction

Exploratory factor analysis has been widely used in social sciences and beyond to measure unobserved constructs of interest, such as ability, attitudes, and behaviors, and for dimensionality reduction. It has been noted early on in the factor analysis literature, particularly with the development of the more precise computational frameworks for maximum likelihood (ML) estimation in Jöreskog (1967) and Jöreskog and Lawley (1968), that the estimation of factor analysis models often results in improper solutions. Such improper solutions involve zero or negative estimates for error variances, and often correlation estimates greater than one in absolute value. Such occurrences are typically referred to as Heywood cases (Heywood, 1931). Martin and McDonald (1975) distinguish between an exact Heywood case in which at least one of the estimates of the error variances is zero but none are negative, and an ultra-Heywood case, in which at least one estimate of the error variances is negative. A zero error variance implies that there is no measurement error, and the factors fully explain the observed variable. That is rare in real applications but, at the same time, does not pose as much

© The Author(s), 2026. Published by Cambridge University Press on behalf of Psychometric Society.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psy.2026.10092 Published online by Cambridge University Press

concern as negative estimates for error variances do. Causes of Heywood cases that have been reported in the literature are model and data dependent and include outliers, non-convergence of associated optimization procedures, under-identification, model misspecification, missing data, and sampling fluctuations combined with a true value close to the boundary for the parameter, small sample sizes, poorly defined factors, and factor over-extraction (see, e.g., Chen et al., 2001; Cooperman & Waller, 2022; Dillon et al., 1987; Kano, 1998; van Driel, 1978; and the references therein). Cooperman and Waller (2022) provide an up-to-date review of the causes, effects, and solutions to Heywood cases in confirmatory and exploratory factor analysis.

The presence of Heywood cases in factor analysis has practical implications. It can produce parameter estimates, standard errors, factor scores, and goodness-of-fit test statistics that cannot be trusted. Cooperman and Waller (2022) found, through a simulation study, that Heywood cases increase the standard errors of factor loadings and bias the factor scores upward. Eliminating items that correspond to estimates which display a Heywood case often moves the Heywood problem to one of the remaining items.

An approach to handle Heywood cases, especially when they are suspected to be due to sampling fluctuations, is by restricting the estimates of the error variances to $[0, \infty)$ either explicitly or by setting negative estimates to zero (see Gerbing & Anderson, 1987 for a discussion). However, this violates regularity conditions of ML estimation, leading to estimators and testing procedures with properties that are hard to evaluate. Another common approach is to impose priors on the loadings, error variances, or both to avoid improper solutions. Estimation, then, proceeds either using a likelihood-based approach with the prior information incorporated via a penalty term (see, e.g., Akaike, 1987; Hirose et al., 2011; Lee, 1981; Martin & McDonald, 1975) or by posterior sampling through MCMC (see, e.g., Lee & Song, 2002). For example, Martin and McDonald (1975) proposed a Bayesian estimation framework in which they maximize not the likelihood but the posterior density, using a prior distribution for the error covariance matrix that assigns zero probability to negative values. They assume a prior distribution that is almost uniform, except that it decreases to zero at the point where an error variance is equal to zero. Lee (1981) also investigated the form of the posterior density under different informative prior distributions, some of which have been designed to deal with Heywood cases. Akaike (1987), in the process of developing a model selection criterion for factor analysis, also encountered the problem of improper solutions and proposed a standard spherical prior distribution of factor loadings and a uniform distribution for the error variances. Hirose et al. (2011) build on Akaike's work by imposing a prior distribution only on the error variances, where the inverse of the diagonal elements of the error covariance matrix has exponential distributions. To the best of our knowledge, there has been no formal proof that such penalties prevent Heywood cases. Furthermore, naive penalization can introduce considerable finite sample bias in the estimation of the factor loadings and error variances, as illustrated in the simulation studies of Section 7.

This article introduces a maximum softly penalized likelihood (MSPL) framework for factor models. Specifically, we derive sufficient conditions on an arbitrary penalty to the log-likelihood function that guarantee that maximum penalized likelihood (MPL) estimation never results in the occurrence of Heywood cases. Furthermore, we show that the penalties proposed in Akaike (1987) and Hirose et al. (2011) satisfy those conditions, while guaranteeing key equivariance properties for factor analysis, namely, equivariance under arbitrary scaling of the data and under factor rotations. To our knowledge, this is the first proof that those two penalties can effectively deal with Heywood cases. We, then, present conditions, under which MPL estimators have the desirable asymptotic properties expected from the ML estimator, namely, consistency and asymptotic normality, under the assumption that the limit of the estimator of the variance–covariance matrix decomposes in an exploratory factor analysis fashion. In particular, we show that this is achieved by requiring that the penalization is soft enough, in a way that adapts to how information about the model parameters accumulates. We also discuss how the Akaike (1987) and Hirose et al. (2011) penalties can be adapted for soft penalization. Although our asymptotic results assume correct specification, in which case Heywood cases are sampling artifacts, our simulation studies about model selection performance in Section 7 suggest that under factor number

https://doi.org/10.1017/psy.2026.10092 Published online by Cambridge University Press

misspecification, MSPL estimation continues to exhibit the least-false behavior expected by the ML estimator (see, e.g., White, 1982).

The remainder of this article is organized as follows. Section 2 briefly presents the factor analysis model. The proposed MPL framework is introduced in Section 3. Section 4 states our existence result of MPL estimates, which rules out the occurrence of Heywood cases, and Section 5 provides the asymptotic behavior of MPL estimators. Section 6 discusses the scaling factors for our MSPL estimators, and Section 7 provides a series of simulation studies that investigate the finite sample performance of MSPL-based estimation and inference and compares them to existing Bayesian approaches. Section 8 gives real data examples and final remarks are provided in Section 9. Proofs of all theoretical results and additional materials are provided in the Supplementary Material.

2. Exploratory factor analysis

The factor analysis model for a random vector of observed variables $\boldsymbol{x}$ and q factors $(q < p)$ is

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{z} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top \in \mathfrak{R}^p$, $\boldsymbol{\Lambda}$ is a $p \times q$ real matrix of factor loadings, $\boldsymbol{z} \sim \mathrm{N}(0_q, \boldsymbol{I}_q)$, $\boldsymbol{\varepsilon} \sim \mathrm{N}(0_p, \boldsymbol{\Psi})$, and $\boldsymbol{z}$ is independent of $\boldsymbol{\varepsilon}$. In the latter expressions, $\boldsymbol{\Psi}$ is a $p \times p$ diagonal matrix with jth diagonal element $\psi_j > 0$, and 0_q is a vector of q zeros, and $\boldsymbol{I}_q$ is the $q \times q$ identity matrix. It follows that $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}$ and $\mathrm{var}(\boldsymbol{x}) = \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}$. The exploratory factor analysis model is identifiable only up to orthogonal rotations of the factor loadings matrix $\boldsymbol{\Lambda}$. Bartholomew et al. (2011, Chapter 3) discuss approaches that resolve unidentifiability.

In the presence of realizations of n independent random vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, the log-likelihood function about the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of the exploratory factor analysis model is

$$C - \frac{n}{2}\left[\log\det(\boldsymbol{\Sigma}) + \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\right) + \sum_{i=1}^{n}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})\right], \tag{2.2}$$

where $C = -np\log(2\pi)/2$, $\bar{\boldsymbol{x}} = \sum_{i=1}^{n}\boldsymbol{x}_i/n$ and $\boldsymbol{S} = \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top/n$ is the sample covariance matrix, assumed to be full rank. Clearly, the maximizer of (2.2) with respect to $\boldsymbol{\mu}$ is $\bar{\boldsymbol{x}}$, and at that point the quadratic term in (2.2) involving $\bar{\boldsymbol{x}}$ and $\boldsymbol{\mu}$ vanishes. Then, the profile log-likelihood about $\boldsymbol{\Psi}$ and $\boldsymbol{\Lambda}$ is

$$\ell(\boldsymbol{\theta}; \boldsymbol{S}) = C - \frac{n}{2}\left[\log\det\left(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}\right) + \mathrm{tr}\left\{\left(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}\right)^{-1}\boldsymbol{S}\right\}\right] \tag{2.3}$$

with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{p(q+1)})^\top = (\lambda_{11}, \ldots, \lambda_{pq}, \psi_1, \ldots, \psi_p)^\top$, where λ_{jk} and ψ_j are the (j,k)th and (j,j)th elements of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$, respectively, $(j = 1, \ldots, p; k = 1, \ldots, q)$. Heywood cases correspond to directions $\{\boldsymbol{\theta}(r)\}_{r\in\mathbb{N}}$ such that the value of $\ell(\boldsymbol{\theta}(r); \boldsymbol{S})$ increases but $\lim_{r\to\infty}\boldsymbol{\Psi}(\boldsymbol{\theta}(r))$ is no longer positive definite. Ultra-Heywood cases, that is, maxima of (2.3) with some of the variances negative, can, of course, be prevented by maximizing the log-likelihood under the constraint that $\psi_j > 0$. Nevertheless, this does not eliminate the possibility of at least one of the ML estimates of $\psi_{11}, \ldots, \psi_{pp}$ being exactly zero.

3. Maximum penalized likelihood for handling Heywood cases

A straightforward way to avoid Heywood cases is to employ an MPL estimator

$$\tilde{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell^*(\boldsymbol{\theta}; \boldsymbol{S}), \tag{3.1}$$

where $\ell^*(\boldsymbol{\theta}; \boldsymbol{S}) = \ell(\boldsymbol{\theta}; \boldsymbol{S}) + P^*(\boldsymbol{\theta}; \boldsymbol{S})$ and $\boldsymbol{\Theta} = \left\{\boldsymbol{\theta} \in \mathfrak{R}^{p(q+1)} : \theta_m > 0, m > pq\right\}$, with a penalty function $P^*(\boldsymbol{\theta})$ that discourages estimates of ψ_{ii} being zero, so that the set

$$\arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell^*(\boldsymbol{\theta}; \boldsymbol{S}) = \left\{\tilde{\boldsymbol{\theta}} \in \boldsymbol{\Theta} : \ell^*(\tilde{\boldsymbol{\theta}}; \boldsymbol{S}) = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell^*(\boldsymbol{\theta}; \boldsymbol{S})\right\},$$

https://doi.org/10.1017/psy.2026.10092 Published online by Cambridge University Press

is nonempty. Toward constructing an MPL estimator that always exists, Akaike (1987) and Hirose et al. (2011) proposed the penalties

$$P^*(\boldsymbol{\theta}) = -\frac{\rho n}{2}\operatorname{tr}\left(\boldsymbol{\Psi}^{-1/2}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1/2}\right), \quad \text{and} \quad P^*(\boldsymbol{\theta}) = -\frac{\rho n}{2}\operatorname{tr}\left(\boldsymbol{\Psi}^{-1/2}\boldsymbol{S}\boldsymbol{\Psi}^{-1/2}\right), \tag{3.2}$$

respectively, for $\rho > 0$. The penalties in (3.2) are attractive because the MPL estimates preserve two particular equivariance properties that the ML estimator has, namely, equivariance under rescaling of the response vectors and equivariance under rotations of the factor loadings. The former is desirable to justify the common practice in factor analysis of setting $\boldsymbol{S}$ in (2.3) to the sample correlation matrix, and the latter is desirable because it ensures that any post-fit rotation of the factors is still the ML estimate of the rotated factors.

To see the equivariance under rescaling of the response vectors, let $\dot{\boldsymbol{x}}_i = \boldsymbol{L}\boldsymbol{x}_i$, for a known, diagonal, invertible $p \times p$ matrix $\boldsymbol{L}$. Then, $\dot{\boldsymbol{\Sigma}} = \operatorname{var}(\dot{\boldsymbol{x}}_i) = \dot{\boldsymbol{\Lambda}}\dot{\boldsymbol{\Lambda}}^\top + \dot{\boldsymbol{\Psi}}$, with $\dot{\boldsymbol{\Lambda}} = \boldsymbol{L}\boldsymbol{\Lambda}$ and $\dot{\boldsymbol{\Psi}} = \boldsymbol{L}\boldsymbol{\Psi}\boldsymbol{L}^\top$, and the sample variance–covariance matrix based on $\dot{\boldsymbol{x}}_1, \ldots, \dot{\boldsymbol{x}}_n$ is $\dot{\boldsymbol{S}} = \boldsymbol{L}\boldsymbol{S}\boldsymbol{L}^\top$. Denoting $\dot{\boldsymbol{\theta}} = (\dot{\lambda}_{11}, \ldots, \dot{\lambda}_{pq}, \dot{\psi}_1, \ldots, \dot{\psi}_p)^\top$, the cyclic property of the trace operator and properties of the determinant for products of invertible matrices can be used to show that $\ell(\dot{\boldsymbol{\theta}}; \dot{\boldsymbol{S}}) = \ell(\boldsymbol{\theta}; \boldsymbol{S}) + \dot{c}$, where $\dot{c}$ does not depend on $\dot{\boldsymbol{\theta}}$. Hence, if $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\Psi}}$ are the maximizers of $\ell(\boldsymbol{\theta}; \boldsymbol{S})$, the maximizers of $\ell(\dot{\boldsymbol{\theta}}; \dot{\boldsymbol{S}})$ are $\boldsymbol{L}\hat{\boldsymbol{\Lambda}}$ and $\boldsymbol{L}\hat{\boldsymbol{\Psi}}\boldsymbol{L}^\top$, respectively. Similar calculations show that, for both penalties in (3.2), $P^*(\dot{\boldsymbol{\theta}}) = P^*(\boldsymbol{\theta}) + \dot{d}$ for a known constant $\dot{d}$ that does not depend on $\dot{\boldsymbol{\theta}}$. Hence, if $\tilde{\boldsymbol{\Lambda}}$ and $\tilde{\boldsymbol{\Psi}}$ are the maximizers of $\ell^*(\boldsymbol{\theta}; \boldsymbol{S})$, the maximizers of $\ell^*(\dot{\boldsymbol{\theta}}; \dot{\boldsymbol{S}})$ are $\boldsymbol{L}\tilde{\boldsymbol{\Lambda}}$ and $\boldsymbol{L}\tilde{\boldsymbol{\Psi}}\boldsymbol{L}^\top$, respectively. The equivariance under rotations of the factors is a direct consequence of the invariance of both $\ell(\boldsymbol{\theta}; \boldsymbol{S})$ and the penalties in (3.2), when $\boldsymbol{\Lambda}$ is replaced by $\boldsymbol{\Lambda}\boldsymbol{Q}$, for an orthogonal $q \times q$ matrix $\boldsymbol{Q}$.

Despite the above attractive equivariance properties, to our knowledge, there has been no formal proof that penalties (3.2) resolve Heywood cases. Furthermore, naive choice of ρ can introduce considerable finite-sample bias in the estimation of $\boldsymbol{\theta}$, as it is illustrated later in the simulations of Section 7.

4. Existence of maximum penalized likelihood estimates

Theorem 4.1 provides general conditions that ensure the existence of MPL estimates and uses them to examine the properties of the penalties (3.2).

Theorem 4.1 (Existence of MPL estimates in factor analysis). *Let* $\boldsymbol{\Theta} = \left\{\boldsymbol{\theta} \in \Re^{p(q+1)} : \theta_m > 0, m > pq\right\}$ *and* $\partial\boldsymbol{\Theta} = \left\{\boldsymbol{\theta} \in \Re^{p(q+1)} : \exists m > pq, \theta_m = 0\right\}$ *and* $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{\Lambda}(\boldsymbol{\theta})^\top + \boldsymbol{\Psi}(\boldsymbol{\theta})$. *Assume that* $\boldsymbol{S}$ *has full rank and that the penalty function* $P^*(\boldsymbol{\theta}) : \boldsymbol{\Theta} \to \Re$ *:*

- (E1) *is continuous on* $\boldsymbol{\Theta}$*;*
- (E2) *is bounded from above on* $\boldsymbol{\Theta}$*, that is,* $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} P^*(\boldsymbol{\theta}) < \infty$*;*
- (E3) *diverges to* $-\infty$ *for any sequence* $\{\boldsymbol{\theta}(r)\}_{r\in\mathbb{N}}$ *such that* $\lim_{r\to\infty}\boldsymbol{\theta}(r) \in \partial\boldsymbol{\Theta}$ *and* $\lim_{r\to\infty}\lambda_{\min}(\boldsymbol{\Sigma}(\boldsymbol{\theta}(r))) > 0$*, where* $\lambda_{\min}(\boldsymbol{A})$ *is the minimum eigenvalue of a matrix* $\boldsymbol{A}$*.*

Then, the set of MPL estimates $\arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell^*(\boldsymbol{\theta}; \boldsymbol{S})$ *is non-empty.*

The proof of Theorem 4.1 is in Section S2.2 of the Supplementary Material. The theorem is model-agnostic in that it only requires $\boldsymbol{S}$ to have full rank and that parameter estimation is conducted using the penalized log-likelihood function of (3.1); no assumptions about the true data generating process are made. Theorem 4.1 establishes that under conditions E1–E3 for the penalty to the log-likelihood, MPL estimation always results in estimates that are not Heywood cases, in the sense that $\tilde{\boldsymbol{\theta}}$ has $\tilde{\psi}_j > 0 (j = 1, \ldots, p)$.

https://doi.org/10.1017/psy.2026.10092 Published online by Cambridge University Press

The penalties by Akaike (1987) and Hirose et al. (2011) in (3.2) satisfy assumptions E1–E3 for $\rho > 0$, and, hence, MPL estimation using either of those results in no Heywood cases. To see that, note that matrix inversion, matrix multiplication, and trace are all continuous operations on $\boldsymbol{\Theta}$. As a result, the penalties in (3.2) are continuous and assumption E1 is satisfied. The penalties in (3.2) can be re-expressed as

$$P^*(\boldsymbol{\theta}) = -\frac{\rho n}{2} \sum_{j=1}^{p} \frac{\boldsymbol{A}_{jj}(\boldsymbol{\theta})}{\boldsymbol{\Psi}_{jj}(\boldsymbol{\theta})}, \tag{4.1}$$

where $\boldsymbol{A}_{jj}(\boldsymbol{\theta}) = \boldsymbol{S}_{jj}$ for the Hirose et al. (2011) penalty, and $\boldsymbol{A}_{jj}(\boldsymbol{\theta}) = \boldsymbol{\lambda}_j(\boldsymbol{\theta})^\top \boldsymbol{\lambda}_j(\boldsymbol{\theta})$ for the Akaike (1987) penalty, where $\boldsymbol{\lambda}_j(\boldsymbol{\theta})$ is the jth row of $\boldsymbol{\Lambda}(\boldsymbol{\theta})$, and $\boldsymbol{C}_{jk}$ denotes the (j,k)th element of the matrix $\boldsymbol{C}$. Note that $\boldsymbol{A}_{jj}(\boldsymbol{\theta})/\boldsymbol{\Psi}_{jj}(\boldsymbol{\theta}) \geq 0$ for both penalties. Hence, (4.1) is bounded above by zero for $\rho > 0$, and E2 is satisfied. Now, consider a sequence $\{\boldsymbol{\theta}(r)\}_{r\in\mathbb{N}}$ such that $\lim_{r\to\infty} \boldsymbol{\theta}(r) \in \partial\boldsymbol{\Theta}$ and $\lim_{r\to\infty} \lambda_{\min}(\boldsymbol{\Sigma}(\boldsymbol{\theta}(r))) > 0$. Then, there exists at least one $j \in \{1,\ldots,p\}$ such that $\boldsymbol{\Psi}_{jj}(\boldsymbol{\theta}(r)) \to 0$. For $\boldsymbol{A}(\boldsymbol{\theta}) = \boldsymbol{S}$, the penalty (4.1) diverges to $-\infty$ as $\boldsymbol{\Psi}_{jj}(\boldsymbol{\theta}) \to 0$. For the Akaike (1987) version, $\boldsymbol{\lambda}_j(\boldsymbol{\theta}(r))^\top \boldsymbol{\lambda}_j(\boldsymbol{\theta}(r))$ can either diverge to ∞ or converge to a constant $c_j > 0$. Only the former can happen for the chosen sequence $\{\boldsymbol{\theta}(r)\}_{r\in\mathbb{N}}$, because, for the latter, $\boldsymbol{\lambda}_j(\boldsymbol{\theta}(r))^\top \boldsymbol{\lambda}_j(\boldsymbol{\theta}(r))$ would need to converge to zero at an appropriate rate, in which case $\boldsymbol{\lambda}_j(\boldsymbol{\theta}(r))^\top \boldsymbol{\lambda}_j(\boldsymbol{\theta}(r)) + \boldsymbol{\Psi}_{jj}(\boldsymbol{\theta})$ converges to zero, resulting in $\boldsymbol{\Sigma}(\boldsymbol{\theta}(r))$ having at least one zero eigenvalue. Hence, E3 is satisfied for both the Akaike (1987) and Hirose et al. (2011) penalties.

Theorem A2 in the Supplementary Material provides an existence result under more general parameterizations of the factor analysis model, which is used for proving the consistency results in Section 5.2, and which might be useful if one wishes to impose further restrictions on the structure of $\boldsymbol{\Sigma}$, as is being done, for example, in confirmatory factor analysis (see, e.g., Bartholomew et al., 2011, Chapter 8).

5. Asymptotics for maximum penalized likelihood

5.1. Consistency

To discuss the consistency of estimates for $\boldsymbol{\Lambda}, \boldsymbol{\Psi}$ in factor analysis models, we must i) define the estimands $\boldsymbol{\Lambda}_0, \boldsymbol{\Psi}_0$, and $\boldsymbol{\Sigma}_0 = \boldsymbol{\Lambda}_0\boldsymbol{\Lambda}_0^\top + \boldsymbol{\Psi}_0$ and ii) ensure identifiability.

If the modeling assumption of Section 2 is met for matrices $\boldsymbol{\Lambda}_0$ and $\boldsymbol{\Psi}_0$, then the latter are the parameter values that identify the data-generating process. This is the viewpoint taken in Kano (1983) in their consistency proofs, where they introduce and use the concept of strong identifiability.

Let $\boldsymbol{B}$ be any $p \times q$ matrix and $\boldsymbol{V}$ be any positive-definite $p \times p$ diagonal matrix and define $\boldsymbol{\Sigma} = \boldsymbol{B}\boldsymbol{B}^\top + \boldsymbol{V}$. A factor model is said to be strongly identifiable if and only if, for any $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$\|\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}\| < \delta \implies \|\boldsymbol{\Lambda}_0 - \boldsymbol{B}\boldsymbol{Q}\| < \varepsilon \quad \text{and} \quad \|\boldsymbol{\Psi}_0 - \boldsymbol{V}\| < \varepsilon \tag{5.1}$$

for some orthogonal matrix $\boldsymbol{Q}$ of order q and where $\|\cdot\|$ is some matrix norm. Since our results are derived in a fixed p and fixed q asymptotic regime, the particular choice of matrix and vector norm is irrelevant for the developments (see Horn & Johnson, 2012, Corollary 5.4.5 and Section S2.1 of the Supplementary Material for details). Hence, throughout, let $\|\cdot\|$ and $\|\cdot\|$ denote any vector and matrix norm, respectively.

Theorem 5.1. *Assume that*

(C1) *the factor model is strongly identifiable;*
(C2) *the set of MPL estimates* $\arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell^*(\boldsymbol{\theta}; \boldsymbol{S})$ *is non-empty;*
(C3) $P^*(\boldsymbol{\theta}) \leq 0$ *for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Then, for any $\varepsilon > 0$, *there exists a* $\delta > 0$ *such that*

$$\|\boldsymbol{S} - \boldsymbol{\Sigma}_0\| < \delta \quad \textit{and} \quad |n^{-1}P^*(\boldsymbol{\theta}_0)| < \delta \implies \|\boldsymbol{\Lambda}_0 - \boldsymbol{\Lambda}(\tilde{\boldsymbol{\theta}})\boldsymbol{Q}\| < \varepsilon \quad \textit{and} \quad \|\boldsymbol{\Psi}_0 - \boldsymbol{\Psi}(\tilde{\boldsymbol{\theta}})\| < \varepsilon \tag{5.2}$$

for some orthogonal $q \times q$ *matrix* $\boldsymbol{Q}$.

https://doi.org/10.1017/psy.2026.10092 Published online by Cambridge University Press

The proof of Theorem 5.1 is in Section S2.3 of the Supplementary Material. Theorem 5.1 shows that if $\boldsymbol{S} \to \boldsymbol{\Sigma}_0$ and $n^{-1}P^*(\boldsymbol{\theta}_0) \to 0$ either in probability or almost surely, then the MPL estimates $\boldsymbol{\Lambda}(\tilde{\boldsymbol{\theta}})$, $\boldsymbol{\Psi}(\tilde{\boldsymbol{\theta}})$ converge to $\boldsymbol{\Lambda}_0$, $\boldsymbol{\Psi}_0$, respectively, in probability or almost surely, up to orthogonal rotations of $\boldsymbol{\Lambda}(\tilde{\boldsymbol{\theta}})$. Note that the conditions that we require of the penalty function are mild; $P^*(\boldsymbol{\theta})$ can be deterministic or depend on the responses, as long as it is pointwise $o_p(n)$.

5.2. $\sqrt{n}$-consistency and asymptotic distribution

Results on the rate of consistency and the asymptotic distribution of the MPL estimator can be derived under a stronger condition on the order of the penalty than that of Theorem 5.1. In particular, we are interested in $\sqrt{n}$-consistency of the MPL estimates, that is,

$$\sqrt{n}\left\|\boldsymbol{\Lambda}(\tilde{\boldsymbol{\theta}})\boldsymbol{Q}_1 - \boldsymbol{\Lambda}(\hat{\boldsymbol{\theta}})\boldsymbol{Q}_2\right\| = o_p(1), \quad \sqrt{n}\left\|\boldsymbol{\Psi}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{\Psi}(\hat{\boldsymbol{\theta}})\right\| = o_p(1),$$

where $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$ denote the MPL and ML estimates, respectively, and $\boldsymbol{Q}_1, \boldsymbol{Q}_2$ are appropriate sequences of orthogonal rotation matrices. Under stronger identification conditions, we can also establish results on the asymptotic distribution of $\tilde{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}$.

Central to such results is the local identification restriction that the Jacobian of $\mathrm{vec}(\boldsymbol{\Sigma}(\boldsymbol{\theta}))$ has full column rank at the parameter of interest $\boldsymbol{\theta}_0$. If the map $\boldsymbol{\theta} \mapsto \mathrm{vec}(\boldsymbol{\Sigma}(\boldsymbol{\theta}))$ is continuously differentiable with full column rank Jacobian at $\boldsymbol{\theta}_0$, then there exists an open neighborhood $\mathcal{U}_0$ around $\boldsymbol{\theta}_0$, such that for any $\varepsilon > 0$, there exists a $\delta > 0$, with

$$\left\|\boldsymbol{\Sigma}(\boldsymbol{\theta}) - \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\right\| < \delta \implies \left\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\right\| < \varepsilon$$

for all $\boldsymbol{\theta} \in \mathcal{U}_0$. This assumption ensures that the information matrix is invertible, which is required for $\sqrt{n}$-asymptotics. The full column rank condition N2 in Theorem 5.2 is, for example, also present in Anderson and Amemiya (1988, Theorems 2 and 3), who establish the asymptotic normal distribution of the $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ in factor analysis models under that and additional conditions.

In the unrestricted model, where $\boldsymbol{\Theta} = \{\boldsymbol{\theta} \in \mathfrak{R}^{p(q+1)} : \theta_m > 0, m > pq\}$, this full rank condition does not hold due to the invariance of the variance–covariance matrix under orthogonal rotations of $\boldsymbol{\Lambda}$. Thus, henceforth, we focus on a restricted parameter space $\tilde{\boldsymbol{\Theta}} \subseteq \mathfrak{R}^d$ with covariance mapping $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{\Lambda}(\boldsymbol{\theta})^\top + \boldsymbol{\Psi}(\boldsymbol{\theta})$ that does not allow for a rank-deficient Jacobian. We further assume that $\tilde{\boldsymbol{\Theta}}$ is contained in $\boldsymbol{\Theta}$ in the sense that for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, there exists a $\tilde{\boldsymbol{\theta}} \in \tilde{\boldsymbol{\Theta}}$ and an orthogonal rotation $\boldsymbol{Q}$, such that $\boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{Q} = \boldsymbol{\Lambda}(\tilde{\boldsymbol{\theta}})$ and $\boldsymbol{\Psi}(\boldsymbol{\theta}) = \boldsymbol{\Psi}(\tilde{\boldsymbol{\theta}})$. Common parameter restrictions encountered in practice, that accommodate this requirement include requiring $\boldsymbol{\Lambda}$ to be upper-triangular, or requiring $\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}$ to be diagonal (see, e.g., Bartholomew et al., 2011, Chapter 3).

Note that the parameter vector defining the matrices $\boldsymbol{\Lambda}_0$ and $\boldsymbol{\Psi}_0$ need not itself be in $\tilde{\boldsymbol{\Theta}}$. Rather, we assume that $\tilde{\boldsymbol{\Theta}}$ contains a parameter vector that corresponds to a loading matrix that can be appropriately rotated to give $\boldsymbol{\Lambda}_0$. Unless $\boldsymbol{\Lambda}_0$ and $\boldsymbol{\Psi}_0$ correspond to a parameter vector in $\tilde{\boldsymbol{\Theta}}$, convergence of loading matrices is therefore naturally stated after appropriate rotation.

Theorem 5.2. *Assume that*

- (N1) *there exists a* $\boldsymbol{\theta}_0 \in \tilde{\boldsymbol{\Theta}}$ *such that* $\boldsymbol{S} \xrightarrow{p} \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$ *as* $n \to \infty$*;*
- (N2) $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$ *is strongly identifiable in* $\tilde{\boldsymbol{\Theta}}$ *and the Jacobian of* $vec(\boldsymbol{\Sigma}(\boldsymbol{\theta}))$ *with respect to* $\boldsymbol{\theta}$ *has full column rank at* $\boldsymbol{\theta}_0$*;*
- (N3) *the set of MPL estimates* $\arg\max_{\boldsymbol{\theta} \in \tilde{\boldsymbol{\Theta}}} \{\ell^*(\boldsymbol{\theta}; \boldsymbol{S})\}$ *is not empty;*
- (N4) $P^*(\boldsymbol{\theta}) = c_n P(\boldsymbol{\theta})$*, where* $P(\boldsymbol{\theta})$ *is nonpositive, deterministic, invariant under orthogonal rotations of* $\boldsymbol{\Lambda}$*, and continuously differentiable on* $\tilde{\boldsymbol{\Theta}}$*, with* $c_n = o_p(\sqrt{n})$ *positive.*

Then, there exist sequences of orthogonal rotation matrices $\boldsymbol{Q}_1, \boldsymbol{Q}_2$ *such that:*

$$\left\|\boldsymbol{\Lambda}(\tilde{\boldsymbol{\theta}})\boldsymbol{Q}_1 - \boldsymbol{\Lambda}_0\right\| \xrightarrow{p} 0, \quad \left\|\boldsymbol{\Psi}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{\Psi}_0\right\| \xrightarrow{p} 0, \tag{5.3}$$

https://doi.org/10.1017/psy.2026.10092 Published online by Cambridge University Press

and

$$\sqrt{n}\left\|\boldsymbol{\Lambda}(\tilde{\boldsymbol{\theta}})\boldsymbol{Q}_1-\boldsymbol{\Lambda}(\hat{\boldsymbol{\theta}})\boldsymbol{Q}_2\right\|\xrightarrow{p}0,\quad \sqrt{n}\left\|\boldsymbol{\Psi}(\tilde{\boldsymbol{\theta}})-\boldsymbol{\Psi}(\hat{\boldsymbol{\theta}})\right\|\xrightarrow{p}0, \tag{5.4}$$

where $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ denote two sequences of the ML and MPL estimates, respectively.

The proof of Theorem 5.2 is in Section S2.4 of the Supplementary Material. Theorem 5.2 as presented above gives conditions about the consistency of ML and MPL estimators, as well as the rate of convergence to one another. If one is interested in establishing the asymptotic distribution of the MPL estimator $\tilde{\boldsymbol{\theta}}$ based on the limiting distribution of the ML estimator $\hat{\boldsymbol{\theta}}$, one can replace the strong identifiability condition on $\boldsymbol{\Lambda}_0, \boldsymbol{\Psi}_0$ with a more stringent point-identification condition on $\boldsymbol{\theta}_0$:

N2*) for any $\varepsilon > 0$, there exists a $\delta > 0$ such that for all $\boldsymbol{\theta} \in \tilde{\boldsymbol{\Theta}}$, $\|\boldsymbol{\Sigma}(\boldsymbol{\theta}) - \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\| < \delta$ implies that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \varepsilon$ and the Jacobian of $\mathrm{vec}(\boldsymbol{\Sigma}(\boldsymbol{\theta}))$ with respect to $\boldsymbol{\theta}$ has full column rank at $\boldsymbol{\theta}_0$.

In this instance, the conclusion of Theorem 5.2 is that $\tilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ and $\sqrt{n}\|\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}\| = o_p(1)$. This stronger identification condition may be required when one wishes to establish the asymptotic distribution of $\tilde{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}$. Slutsky's lemma then implies that if $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ has a normal distribution asymptotically, then $\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ with a penalty scaled as in N4 has the same asymptotic distribution.

Note that the conditions of Theorems 5.1 and 5.2 only require that the limit of the estimator of the variance–covariance matrix $\boldsymbol{S}$ decomposes in an exploratory factor analysis fashion, and normality of errors is not essential.

6. Maximum softly penalized likelihood

Theorem 4.1 establishes conditions on $P^*(\boldsymbol{\theta})$ that ensure the existence of the MPL estimates. On the other hand, Theorems 5.1 and 5.2 involve sufficient conditions on the order of the penalty $P^*(\boldsymbol{\theta})$ for the consistency of the MPL estimator. Specifically, if $P^*(\boldsymbol{\theta}) = o_p(\sqrt{n})$, the respective order conditions in Theorems 5.1 and 5.2 are satisfied.

Suppose that $P^*(\boldsymbol{\theta}) = c_n P(\boldsymbol{\theta})$, where the functional part $P(\boldsymbol{\theta})$ satisfies the conditions of Theorem 4.1 for the existence of MPL estimates, and where $c_n > 0$ is a scaling factor. One way to derive a suitable scaling factor is to consider how information about the model parameters accumulates as n increases. For example, we can derive a principled heuristic for c_n by considering the exploratory factor analysis model in (2.3) under independence (i.e., $\boldsymbol{\Lambda}$ is a matrix of zeros). The unknown parameters are the vector of variances $\sigma_1^2, \ldots \sigma_p^2$, and the information matrix about those parameters is a diagonal matrix with jth diagonal element $n/(2\sigma_j^4)$. Standard results on the asymptotic distribution of the ML estimator give that $\sqrt{n/2}(s_j^2/\sigma_j^2 - 1)$ converges in distribution to a standard normal random variable for all $j \in \{1, \ldots, p\}$. The rate of information accumulation for each coordinate is $\sqrt{n/2}$ and, hence, we can choose $c_n = \sqrt{2/n}$. This choice is in line with the requirements of Theorems 5.1 and 5.2 for any exploratory factor analysis model, while ensuring that the penalization strength is asymptotically negligible. We call MSPL estimation, MPL estimation with asymptotically negligible penalties that guarantee existence and $\sqrt{n}$-consistency. In Section 4, we showed that the conditions for the existence of the MPL estimates are satisfied for the penalties (3.2) in Akaike (1987) and Hirose et al. (2011). In particular, both penalties have the form $P^*(\boldsymbol{\theta}) = \rho n P(\boldsymbol{\theta})/2$, $\rho > 0$, $P(\boldsymbol{\theta}) \leq 0$, and $P(\boldsymbol{\theta}) = O(1)$. Both penalties can be adapted for MSPL estimation by setting $\rho = 2\sqrt{2/n^3}$.

7. Simulation studies

We conduct a series of simulation experiments to compare the performance of ML estimation, MPL estimation based on the penalties (3.2) with vanilla choices of ρ, and MSPL estimation. The methods are compared in terms of their ability to handle Heywood cases, frequentist properties of the estimators, and selection of the number of factors when Heywood cases are present.

https://doi.org/10.1017/psy.2026.10092 Published online by Cambridge University Press

Table 1. Loading matrix settings A_3 and B_3

Item	Setting A_3			Setting B_3		
1	0.80	0	0	0.80	0	0
2	0.65	0	0	0.80	0	0
3	0.45	0	0	0.80	0	0
4	0	0.80	0	0	0.80	0
5	0	0.65	0	0	0.80	0
6	0	0.45	0	0	0.80	0
7	0	0	0.80	0	0	0.30
8	0	0	0.65	0	0	0.30
9	0	0	0.45	0	0	0.30

Our simulation settings have been informed by the simulation-based results reported in Cooperman and Waller (2022). Specifically, Cooperman and Waller (2022) identified, through an extensive simulation study, the following causes of Heywood cases in order of importance: item-to-factor ratio, model specification (correct, fitting one factor less than the actual number, and fitting one extra factor than the actual number), sample size, and loading matrix pattern (low against high loadings on the same factor, and factors with all low loadings against factors with all high loadings).

We fix the number of factors to $q = 3$. We then consider sample sizes $n \in \{50, 100, 400\}$ and the item-to-factor ratios 3:1, 5:1, and 8:1, and let the loading matrix $\mathbf{\Lambda}$ vary across experimental settings. Specifically, for the loading matrix with item-to-factor ratio 3:1, we use the matrices in settings A_3 and B_3 in Table 1, which are motivated from the settings of Cooperman and Waller (2022, Table 2). Setting A_3 decreases the factor loadings sequentially, while setting B_3 assumes two strong and one weak factor. Setting A_5 for the 5:1 ratio and setting A_8 for the 8:1 ratio are defined correspondingly to A_3, where we choose the nonzero column blocks to be $(0.80, 0.65, 0.50, 0.35, 0.20)$ and $(0.80, 0.70, 0.60, 0.50, 0.40, 0.30, 0.20, 0.10)$, respectively. For settings B_5 for the 5:1 ratio and B_8 for the 8:1 ratio, we simply repeat the non-zero loadings in Table 1 according to the item-to-factor ratio. The specific variances $\mathbf{\Psi}$ are set so that the diagonal elements of $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}$ are all one.

We compare the ML estimator with the MSPL estimators using appropriately scaled versions of the penalties (3.2) with $\rho = 2\sqrt{2/n^3}$ based on the discussion in Section 6. We refer to those penalties as "Akaike$[n^{-1/2}]$" and "Hirose$[n^{-1/2}]$." We also consider MPL with the non-decaying scaling $\rho = 1$, which was also used in Hirose et al. (2011). We refer to those penalties as "Akaike$[n]$" and "Hirose$[n]$."

For each combination of loading matrix, sample size, and item-to-factor ratio, we draw 1,000 independent samples according to the factor analysis model (2.1) and, for each sample, we compute the estimates of (3.1). The estimates are computed by first getting MPL estimates from 100 iterations of an EM-maximization of the penalized log-likelihood, which we then use as starting values for a Newton–Raphson optimization routine. We identify Heywood cases heuristically, when at least one of the following occurs: the estimation procedure fails, the normalized gradient $\{\nabla\nabla^\top \ell^*(\boldsymbol{\theta})\}^{-1}\nabla\ell^*(\boldsymbol{\theta})$ has at least one element with absolute value greater than 10^{-4}, and at least one of the estimates of $\psi_1, \ldots, \psi_p$ is less than 10^{-4}.

Figure 1 shows the percentage of samples that have been identified as Heywood cases. It is evident that ML estimation results in a considerable number of Heywood cases across experimental settings, while, as expected from Theorem 4.1, MPL and MSPL estimation are effective in dealing with them. The negligible fraction of estimates that are identified as Heywood cases for MPL and MSPL estimation are attributable to the heuristics we use for the identification of Heywood cases and can be eliminated by less stringent heuristics.

https://doi.org/10.1017/psy.2026.10092 Published online by Cambridge University Press

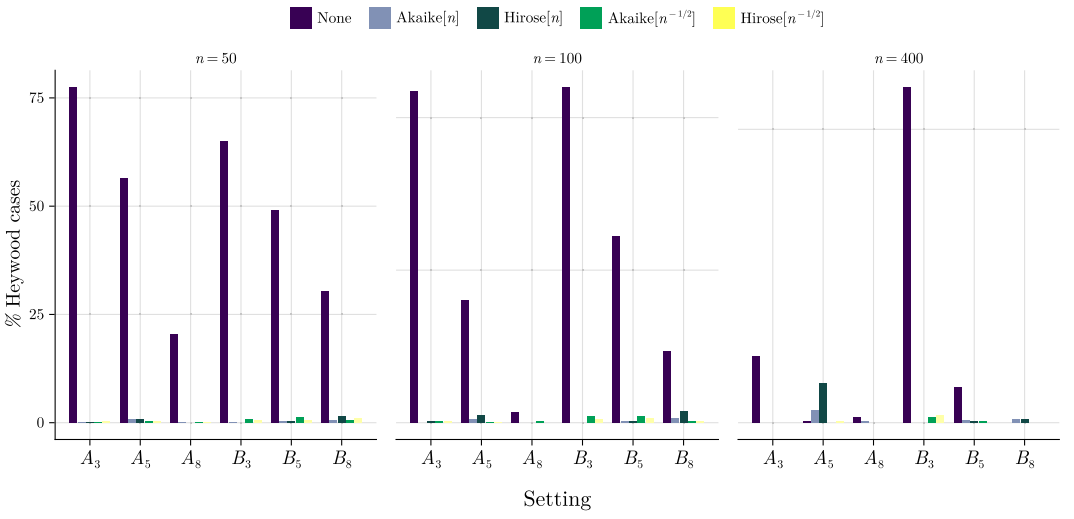


Figure 1. Percentage of samples (out of 1,000) that have been identified as Heywood cases for ML ("None"), MPL with Akaike[n] and Hirose[n] penalties, and MSPL with Akaike[$n^{-1/2}$] and Hirose[$n^{-1/2}$] penalties, $n \in \{50, 100, 400\}$, and loading matrix settings A_3, B_3, A_5, B_5, A_8, and B_8.

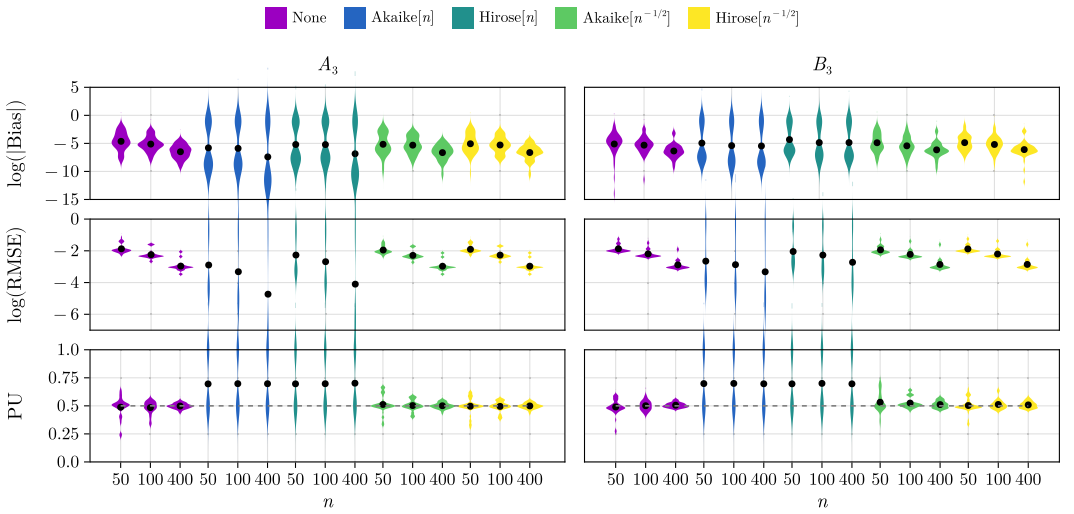


Figure 2. Violin plots of estimates of log(|Bias|) (top panel), log(RMSE) (middle panel), and probability of underestimation (bottom panel) for the elements of $\Lambda\Lambda^\top$, for each estimator, $n \in \{50, 100, 400\}$, and loading matrix settings A_3 and B_3. The average over all elements for each setting is noted with a dot.

We evaluate the finite-sample performance of ML, MPL, and MSPL estimators in terms of bias, probability of underestimation, and root mean-squared error (RMSE), estimated excluding the samples that have been identified as Heywood cases.

Figure 2 shows violin plots of coordinatewise estimates of the logarithm of the absolute bias, the logarithm of the RMSE, and the probability of underestimation of the unique elements of $\boldsymbol{\Lambda\Lambda}^\top$, for each estimator, $n \in \{50, 100, 400\}$, and loading matrix settings A_3 and B_3. A black dot indicates the average over all coordinates in each specification. The MSPL estimates, with soft scaling of order $n^{-1/2}$, exhibit the smallest or close to the smallest bias across all methods and a rate of decay that is in line with Theorem 5.2. Notably, the Akaike[n] and Hirose[n] penalized estimators exhibit large finite sample bias.

https://doi.org/10.1017/psy.2026.10092 Published online by Cambridge University Press

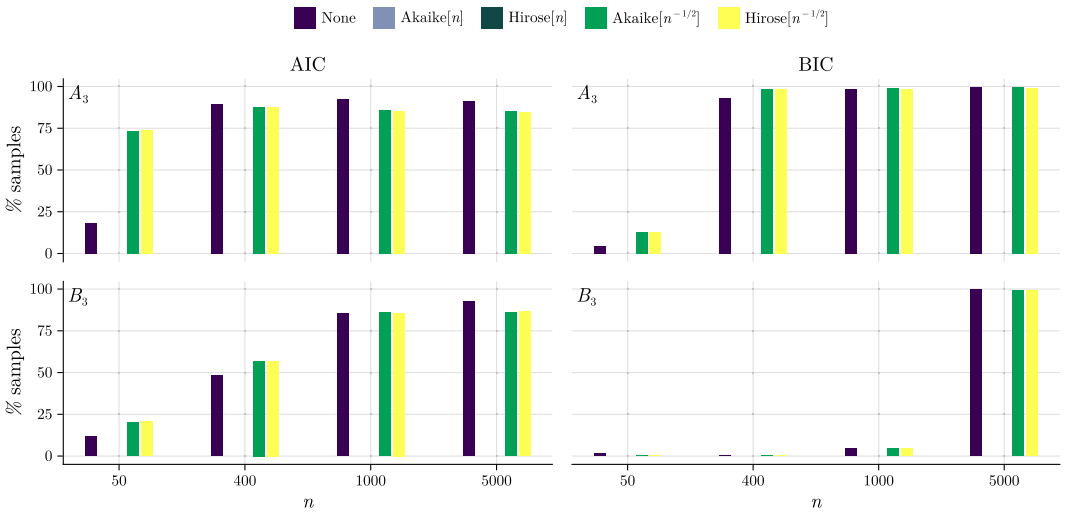


Figure 3. Percentage of times the model with *three* factors is selected for each estimator, $n \in \{50, 400, 1000, 5000\}$, and loading matrix settings A_3 and B_3, using AIC and BIC. The absence of vertical bars pertaining to the Akaike[n]- and Hirose[n]-based model selection procedures indicate that these methods have never selected the correct model.

We also see that the MSPL estimators are well calibrated, with a probability of underestimation close to 1/2 across all settings. In contrast, the Akaike[n] and Hirose[n] estimators consistently underestimate the elements of $\mathbf{\Lambda}\mathbf{\Lambda}^\top$. This underestimation is expected from the excessive penalization that results from using a scaling factor of order n. Similarly to bias, the RMSE of the MSPL estimates is the lowest or close to the minimal RMSE across all methods. The coordinatewise estimates of the logarithm of absolute bias, the logarithm of RMSE, and probability of underestimation for loading matrix settings A_5 and B_5, and A_8 and B_8 are shown in Figures S1 and S2 in the Supplementary Material, respectively. The findings are the same as above for A_3 and B_3.

We also assess the performance of AIC and BIC selection of the number of factors based on each estimator (see Akaike, 1987 and Hirose et al., 2011 for details on these criteria for factor models) for item-to-factor ratio 3:1 with settings A_3 and B_3 for the loading matrix, and $n \in \{50,400,1,000,5,000\}$. We note that when fitting models with $q \in \{1,2\}$, the model is misspecified and the assumptions of Theorems 5.1 and 5.2 do not apply. This design choice mirrors common practitioner behavior in factor selection and motivates our inclusion of this scenario. We fit the factor analysis model (2.1) for $q \in \{1,\ldots,5\}$. The AIC and BIC are computed from the unpenalized log-likelihood evaluated at the ML and MSPL estimates, respectively. We use all the samples, including those that have been identified as leading to Heywood cases, for computing AIC and BIC. Figure 3 shows the percentage of times that the model with three factors was selected with AIC and BIC at each estimator and $n \in \{50,400,1,000,5,000\}$. We note that AIC- and BIC-based model selection performs as expected with the MSPL estimators with Akaike[$n^{-1/2}$] and Hirose[$n^{-1/2}$] penalties. BIC-based model selection selects the correct model with increasing probability as n increases, which is the result of the consistency of BIC-based model selection (see, e.g., Claeskens & Hjort, 2008, Chapter 4). For $n = 50$, model selection based on MSPL estimators is also found to outperform that based on the ML estimator, most probably due to the strong handling of Heywood cases, even in small samples. On the other hand, the MPL estimators with Akaike[n] and Hirose[n] penalties result in poor performance in both AIC- and BIC-based model selection, mainly due to the strength of the penalty.

For $n < 5,000$, we observe the poor performance of BIC in selecting the correct number of factors in setting B_3 across all estimators which we attribute to the two strong and one extremely weak factor (see Table 1). As Table 2 shows, BIC identifies the two strong factors in the majority of cases for small n, and starts identifying the weak factor more frequently slowly as n increases with ML or MSPL estimators.

https://doi.org/10.1017/psy.2026.10092 Published online by Cambridge University Press

Table 2. Percentage of times each number of factors has been selected using minimum BIC, for ML and MSPL with Akaike[n], Hirose[n], Akaike[$n^{-1/2}$], and Hirose[$n^{-1/2}$] penalties, under loading matrix setting B_3 and $n \in \{50, 400, 1000, 5000\}$

n	q	None	Akaike[n]	Hirose[n]	Akaike[$n^{-1/2}$]	Hirose[$n^{-1/2}$]
50	1	5.8	100.0	100.0	1.5	1.5
	2	92.6	0.0	0.0	98.2	98.2
	3	1.5	0.0	0.0	0.3	0.3
	4	0.1	0.0	0.0	0.0	0.0
400	1	0.0	4.7	65.1	0.0	0.0
	2	99.5	95.3	34.9	99.5	99.5
	3	0.5	0.0	0.0	0.5	0.5
	4	0.0	0.0	0.0	0.0	0.0
1,000	1	0.0	0.0	0.0	0.0	0.0
	2	95.2	100.0	100.0	95.3	95.4
	3	4.8	0.0	0.0	4.7	4.6
	4	0.0	0.0	0.0	0.0	0.0
5,000	1	0.0	0.0	0.0	0.0	0.0
	2	0.0	100.0	100.0	0.0	0.0
	3	100.0	0.0	0.0	100.0	100.0
	4	0.0	0.0	0.0	0.0	0.0

8. Real data examples

We estimate the factor model (2.1) using ML, and MSPL using the Akaike[$n^{-1/2}$] and Hirose[$n^{-1/2}$] penalties for three data sets, where Heywood cases have been encountered in published work. The data sets are i) the Davis data (Rao, 1955), which involve $n = 421$ observations and $p = 9$ items, ii) the Emmett data (Emmett, 1949; Lawley & Maxwell, 1971), which involve $n = 211$ observations and $p = 9$ items, and iii) the Maxwell data (Maxwell, 1961, and Lawley & Maxwell, 1971, p. 44), which involve $n = 810$ observations and $p = 10$ items. Heywood cases result in the ML estimates of the factor model (2.1) with $q = 2$ for the Davis data, $q = 5$ for the Emmett data, and $q = 4$ for the Maxwell data. The three data sets have also been analyzed in Akaike (1987).

Table 3 gives the estimates of the communalities $\sum_{k=1}^{q} \lambda_{jk}^2 (j = 1, \ldots, p)$ using ML, and MSPL with Akaike[$n^{-1/2}$] and Hirose[$n^{-1/2}$] penalties, across different number of factors, along with the corresponding AIC and BIC values. As expected, ML estimation can lead to Heywood cases, which manifest as atypically large estimated communalities. In contrast, and as expected, there are no Heywood cases when MSPL estimation is used, and communality estimates are reasonable with no substantial impact on AIC and BIC values. Specifically, in the Davis data set, item 1 has an atypically large communality ML estimate for $q = 2$, while the MSPL estimates are all within reasonable ranges. MSPL estimation also resolves the Heywood cases that result in atypically large ML communality estimates for the Emmett data for $q = 4$ and $q = 5$, and the Maxwell data for $q = 4$. Tables S1–S3 in the Supplementary Material show the estimates of $\boldsymbol{\Psi}$ and $\boldsymbol{\Lambda}$ for the Emmett data, for $q \in \{1, \ldots, 5\}$. As is apparent, the large ML estimated communalities for $q = 4$ and item 3, and $q = 5$ and item 4, correspond to negative ML variance estimates for those items. In contrast, and as expected all MSPL variance estimates are positive. Notably, due to soft penalization, the ML and MSPL estimates that do not correspond to Heywood cases are similar.

https://doi.org/10.1017/psy.2026.10092 Published online by Cambridge University Press

Table 3. Estimated communalities ($\times 10^3$) for the Davis, Emmett, and Maxwell data, using ML, and MSPL with Akaike[$n^{-1/2}$] and Hirose[$n^{-1/2}$] penalties, for $q \in \{1, \ldots, 5\}$, with AIC and BIC values

		Item											
Method	q	1	2	3	4	5	6	7	8	9	10	AIC	BIC
Davis data													
ML	1	658	661	228	168	454	800	705	434	703		1,694.69	1,767.46
	2	**14,145**	634	224	176	463	813	705	439	702		1,683.02	1,788.12
Akaike[$n^{-1/2}$]	1	654	657	227	167	452	795	701	432	699		1,694.70	1,767.47
	2	877	660	226	185	467	808	698	438	696		1,684.89	1,790.00
Hirose[$n^{-1/2}$]	1	658	661	228	168	454	799	705	434	703		1,694.69	1,767.46
	2	882	664	227	187	470	813	703	441	701		1,684.88	1,789.99
Emmett data													
ML	1	510	537	300	548	390	481	525	224	665		1,176.56	1,236.90
	2	538	536	332	809	592	778	597	256	782		984.56	1,071.71
	3	550	573	383	788	619	823	600	538	769		977.58	1,088.20
	4	544	556	**14,307**	797	612	800	604	737	773		986.51	1,117.23
	5	547	645	376	**8,877**	553	991	681	532	759		992.27	1,139.76
Akaike[$n^{-1/2}$]	1	505	531	297	543	386	476	519	222	658		1,176.57	1,236.91
	2	532	528	329	797	585	768	590	253	770		984.59	1,071.74
	3	543	565	379	777	612	809	593	532	758		977.62	1,088.24
	4	540	557	600	789	611	797	597	615	763		986.85	1,117.58
	5	547	632	405	820	653	824	646	492	758		992.74	1,140.23
Hirose[$n^{-1/2}$]	1	510	537	300	548	390	481	524	224	665		1,176.57	1,236.90
	2	538	536	333	807	592	778	597	256	780		984.57	1,071.72
	3	550	573	384	788	620	820	600	536	768		977.60	1,088.22
	4	547	565	605	800	620	807	605	619	773		986.83	1,117.56
	5	555	641	410	833	662	836	654	496	769		992.72	1,140.20
Maxwell data													
ML	1	585	230	567	306	375	159	366	143	198	114	6,442.96	6,536.88
	2	594	250	648	350	442	227	649	359	323	381	5,954.08	6,090.25
	3	631	384	694	356	586	228	664	359	325	378	5,893.17	6,066.92
	4	603	370	702	356	676	199	725	**27,024**	276	391	5,842.66	6,049.27
Akaike[$n^{-1/2}$]	1	584	230	566	306	374	159	365	143	198	114	6,442.96	6,536.88
	2	593	250	647	350	442	226	648	359	322	380	5,954.08	6,090.25
	3	631	384	692	356	584	228	663	359	325	378	5,893.17	6,066.92
	4	619	382	695	363	637	225	707	903	314	405	5,847.36	6,053.98
Hirose[$n^{-1/2}$]	1	584	230	567	306	374	159	366	143	198	114	6,442.96	6,536.88
	2	594	250	648	350	442	227	649	359	323	381	5,954.08	6,090.25
	3	631	384	693	356	585	228	663	359	325	378	5,893.17	6,066.92
	4	620	383	696	364	638	225	708	905	315	406	5,847.36	6,053.98

Note: Heywood cases are shown in bold.

https://doi.org/10.1017/psy.2026.10092 Published online by Cambridge University Press

9. Concluding remarks

In this article, we introduced a novel MSPL framework for factor analysis models to address improper solutions known as Heywood cases that frequently occur in statistical practice. Heywood cases can lead to unstable and inconclusive results related to factor loading estimates and factor scores, as well as inaccurate inferences and model selection. Our approach provides a comprehensive blueprint for constructing penalties and scaling factors that ensure the existence of estimators within the admissible parameter space and avoid the proposed ad hoc solutions in the literature. Our work focuses on exploratory factor analysis, but the proposed estimator can also be applied in a confirmatory factor analysis setting by enforcing additional constraints that practitioners might wish to impose on the model (e.g., zero loadings, equal error variances, etc.).

We provide sufficient conditions for the existence of the MPL estimator in factor analysis, together with the asymptotic properties of consistency and asymptotic normality of the MPL estimators. Additionally, we derive decay rates for the scaling of the penalty function to ensure consistency and asymptotic normality of the MSPL estimators, thus preserving the favorable asymptotic properties expected by the ML estimator. Through extensive simulation studies, we compared MSPL with appropriately scaled versions of the penalties proposed by Hirose et al. (2011), which are derived from Bayesian considerations and thus lack soft penalization by default. The MSPL estimators are found to recover the performance expected from ML theory while resolving the issues related to Heywood cases, across various model specifications, sample sizes, and item-to-factor ratios, making them a valuable tool for practical applications in exploratory and potentially confirmatory factor analysis. Our findings further reveal that naive penalties not only can undermine frequentist properties, in terms of higher bias, RMSE, and probability of underestimation, but can also have a deteriorating effect on the performance of model selection procedures. Our framework enables hypothesis testing by ensuring the existence of interior estimates in finite samples, even when the ML estimator may fail to exist, while preserving standard ML asymptotic behavior.

A limitation of our results is that they assume the probability limit of the sample covariance lies within the exploratory factor analysis model class; extending the theory to a least-false interpretation under covariance misspecification, in the spirit of White (1982), is a natural and important direction for future work. Further research directions include exploring alternative penalty functions, within the MSPL framework, for the factor analysis model (2.1) and for related models such as logistic factor analysis (see Bartholomew et al., 2011, Chapter 4). For example, in the logistic model, steep item characteristic curves lead to infinite estimates of the loadings (see, e.g., Wang & Yotebieng, 2023 for a discussion of Heywood cases in item response models for binary data). MSPL can be readily extended to handle those cases, too.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psy.2026.10092.

Data availability statement. The supplementary material is also available at https://github.com/psterzinger/FAPL, and consists of the three folders "code," "results," "figures," and the Supplementary Material. The latter provides the proofs to our results and evidence from additional simulation studies and numerical examples to those presented in the main text. The "code" directory contains scripts to reproduce the numerical analyses, simulations, graphics, and tables in the main text and the Supplementary Material. The "results" and "figures" directories provide all results and figures from the numerical experiments and analyses in the main text and the Supplementary Material, respectively.

Author contributions. All authors contributed equally to this work.

Funding statement. This research received no specific grant funding from any funding agency, commercial, or not-for-profit sectors.

Competing interests. The authors declare none.

https://doi.org/10.1017/psy.2026.10092 Published online by Cambridge University Press

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*, 317–332. https://doi.org/10.1007/BF02294359

Anderson, T. W., & Amemiya, Y. (1988). The asymptotic normal distribution of estimators in factor analysis under general conditions. *The Annals of Statistics*, *16*(2), 759–771. https://doi.org/10.1214/aos/1176350834

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. (3rd ed.) John Wiley. https://doi.org/10.1002/9781119970583

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research*, *29*(4), 468–508. https://doi.org/10.1177/0049124101029004003

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge University Press.

Cooperman, A. W., & Waller, N. G. (2022). Heywood you go away! Examining causes, effects, and treatments for Heywood cases in exploratory factor analysis. *Psychological Methods*, *27*(2), 156–176. https://doi.org/10.1037/met0000384

Dillon, W. R., Kumar, A., & Mulani, N. (1987). Offending estimates in covariance structure analysis: Comments on the causes of and solutions to Heywood cases. *Psychological Bulletin*, *101*(1), 126–135. https://doi.org/10.1037/0033-2909.101.1.126

Emmett, W. G. (1949). Factor analysis by Lawley's method of maximum likelihood. *British Journal of Psychology, Statistical Section*, *2*, 90–97. https://doi.org/10.1111/j.2044-8317.1949.tb00270.x

Gerbing, D., & Anderson, J. (1987). Improper solutions in the analysis of covariance structures: Their interpretability and a comparison of alternate respecifications. *Psychometrika*, *52*(1), 99–111. https://doi.org/10.1007/BF02293958

Heywood, H. B. (1931). On finite sequences of real numbers. *Proceedings of the Royal Society, A*, *134*, 486–510. https://doi.org/10.1098/rspa.1931.0209

Hirose, K., Kawano, S., Konishi, S., & Ichikawa, M. (2011). Bayesian information criterion and selection of the number of factors in factor analysis models. *Journal of Data Science*, *9*(2), 243–259. https://doi.org/10.6339/JDS.201104˙09(2).0007

Horn, R. A., & Johnson, C. R. (2012). *Matrix analysis*. Cambridge University Press. https://doi.org/10.1017/cbo9781139020411

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, *32*, 443–482. https://doi.org/10.1007/BF02289658

Jöreskog, K. G., & Lawley, D. N. (1968). New methods in maximum likelihood factor analysis. *British Journal of Mathematical and Statistical Psychology*, *21*(1), 85–96. https://doi.org/10.1111/j.2044-8317.1968.tb00399.x

Kano, Y. (1983). Consistency of estimators in factor analysis. *Journal of the Japan Statistical Society, Japanese Issue*, *13*(2), 137–144. https://doi.org/10.11329/jjss1970.13.137

Kano, Y. (1998). Improper solutions in exploratory factor analysis: Causes and treatments. In A. Rizzi, M. Vichi, & H. H. Bock (Eds.), *Advances in data science and classification*, Studies in Classification, Data Analysis, and Knowledge Organization (pp. 375–382). Springer. https://doi.org/10.1007/978-3-642-72253-0˙51

Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method*. (2nd ed.) Butterworth.

Lee, S. Y. (1981). A Bayesian approach to confirmatory factor analysis. *Psychometrika*, *46*, 153–160. https://doi.org/10.1007/BF02293896

Lee, S.-Y., & Song, X.-Y. (2002). Bayesian selection on the number of factors in a factor analysis model. *Behaviometrika*, *29*, 23–39. https://doi.org/10.2333/bhmk.29.23

Martin, J. K., & McDonald, R. P. (1975). Bayesian estimation in unrestricted factor analysis; a treatment for Heywood cases. *Psychometrika*, *40*, 505–517. https://doi.org/10.1007/BF02291552

Maxwell, A. E. (1961). Recent trends in factor analysis. *Journal of the Royal Statistical Society, Series A*, *124*, 49–59. https://doi.org/10.2307/2343153

Rao, C. R. (1955). Estimation and tests of significance in factor analysis. *Psychometrika*, *20*, 93–111. https://doi.org/10.1007/BF02288983

van Driel, O. P. (1978). On various causes of improper solutions in maximum likelihood factor analysis. *Psychometrika*, *43*, 225–243. https://doi.org/10.1007/BF02293865

Wang, S., & Yotebieng, M. (2023). Heywood cases in unidimensional factor models and item response models for binary data. *Applied Psychological Measurement*, *47*(2), 141–154. https://doi.org/10.1177/0146621623115170

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1–25. https://doi.org/10.2307/1912526

Cite this article: Sterzinger, P., Kosmidis, I. and Moustaki, I. (2026). Maximum Softly Penalized Likelihood in Factor Analysis. *Psychometrika*, 494–507. https://doi.org/10.1017/psy.2026.10092

https://doi.org/10.1017/psy.2026.10092 Published online by Cambridge University Press