Statistical Papers
https://doi.org/10.1007/s00362-022-01308-w

REGULAR ARTICLE

Squared error-based shrinkage estimators of discrete probabilities and their application to variable selection

Małgorzata Łazęcka[1,2] · Jan Mielniczuk[1,2]

Received: 11 June 2021 / Revised: 13 January 2022 / Accepted: 16 March 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract

In the paper we consider a new approach to regularize the maximum likelihood estimator of a discrete probability distribution and its application in variable selection. The method relies on choosing a parameter of its convex combination with a low-dimensional target distribution by minimising the squared error (SE) instead of the mean SE (MSE). The choice of an optimal parameter for every sample results in not larger MSE than MSE for James–Stein shrinkage estimator of discrete probability distribution. The introduced parameter is estimated by cross-validation and is shown to perform promisingly for synthetic dependence models. The method is applied to introduce regularized versions of information based variable selection criteria which are investigated in numerical experiments and turn out to work better than commonly used plug-in estimators under several scenarios.

Keywords Discrete distribution · Shrinkage estimator · Stochastic measure of accuracy · Regularisation parameter · Conditional mutual information · Variable selection · Crossvalidation · Squared error · Gini index

1 Introduction

The aim of the present paper is to introduce a new regularisation method of discrete probabilities and to apply it for information based non-parametric variable selection for discrete variables. We begin by shortly describing issues of variable selection

✉ Jan Mielniczuk
jan.mielniczuk@ipipan.waw.pl

Małgorzata Łazęcka
malgorzata.lazecka@ipipan.waw.pl

[1] Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland

[2] Institute of Computer Science, Polish Academy of Sciences Warsaw, Jana Kazimierza 5, 01-248 Warsaw, Poland

Published online: 07 April 2022

and why regularisation of probability estimates is important there. Variable selection methods belong to the most active areas of research in data mining nowadays, especially for high-dimensional data when the number of predictors is large and can be substantially larger than the sample size. Important direction in this area focuses on model-agnostic approaches. This is due to a growing concern about behaviour of model-based methods, which assume specific structure of data generation, when these assumptions are not met. Here, for discrete case, we study a fully non-parametric and model-free approach based on the conditional mutual information (CMI). This has several additional advantages. Namely, CMI based approach detects non-linear dependencies and is able to take redundancies between variables into account. Moreover, it is a versatile method covering classification as well as regression tasks. For high-dimensional data, however, such methods face a problem of inadequate estimation of probability distributions and pertaining information theoretic quantities by fraction estimators, as a number of samples needed to estimate well the conditional probabilities involved grows exponentially with the size of conditioning set. One of methods applied to approach this problem is to use the shrinkage estimators instead of the fraction estimators. The origin of this method is Laplace correction which is used when one or several cells have zero frequencies. The main idea is to take advantage of the opposite characteristics of two estimators: one which has a small bias but a relatively large variance and the second, low-dimensional one, which in contrast, has a small variance and a large bias.

Advantages of information theoretic variable selection lead to a legion of specific methods differing in criterion function, search techniques and stopping rules considered. We refer to Brown et al. (2012) and Vergara and Estevez (2014) for comprehensive reviews. Nowadays, most commonly adapted search technique is sequential forward selection (SFS). At its subsequent step, SFS checks usefulness of all potential candidates to predict binary target Y given set of variables X_S of size $|S|$ chosen from the pool of all available variables $X_1, \ldots, X_p$ and picks the most promising one among them. The most obvious choice of a selection criterion is the CMI between potential candidate X and Y given X_S. However, due to intrinsic estimation difficulties of CMI when $|S|$ is large, its many approximations were proposed which avoid curse of dimensionality problem, starting from a univariate filter, two-dimensional approximations: MIFS (mutual information feature selection Battiti 1994), JMI (joint mutual information Yang and Moody 1999), mRMR (minimal redundancy maximal relevance Peng et al. 2005), CIFE (conditional informative feature extraction Lin and Tang 2006; Kubkowski et al. 2021) and their higher order counterparts (see Vinh et al. 2016; Sechidis et al. 2019; Pawluk et al. 2019). Stopping rules for variable selection criteria are studied e.g. in Mielniczuk and Teisseyre (2019) and Borboudakis and Tsamardinos (2019).

However, even when the approximations of CMI are used as the selection criteria a problem of inadequate estimation of probabilities involved still remains, especially when the sample size is small and the number of possible predictor values is large. This problem has been addressed in the context of entropy and mutual information estimation by Hausser and Strimmer (2007) who used idea of shrinkage (cf. James and Stein 1961). They proposed James–Stein shrinkage estimator of discrete probabilities and compared it to others, e.g. Bayesian methods, which impose the Dirichlet prior on the probability distribution. In Scutari and Brogini (2016) the approach has been used to estimate CMI. Sechidis et al. (2019) considered a variant of this method, more suitable for multivariate vectors having dependent components, with a specific aim of constructing regularised estimators of the information theoretic selection criteria.

In the present paper we extend the methodology of Sechidis et al. (2019). Our contribution is threefold. First, we propose a new way of regularisation, based on (global) squared error (SE) minimisation reminiscent of integrated SE minimisation in density estimation, see e.g. Hall (1982, 1983), Stone (1984) and Scott (2001). The idea is to choose a regularisation parameter in such a way that the resulting convex combination of ML estimator and a chosen low-dimensional target distribution has the smallest global squared distance from *the true unknown* probability mass function. This has immediate theoretical advantages as we show that MSE for the proposed estimator is not larger than MSE for James–Stein shrinkage estimator of discrete probability distribution [see inequality (18)]. We stress that the result is valid for *any* sample size. Secondly, we shed some new light on regularisation parameters introduced in Hausser and Strimmer (2007) and Sechidis et al. (2019) as well as on the one introduced in this paper. Thirdly, we examine usefulness of proposed shrinkage estimators for estimation of probabilities, CMIs as well as for variable selection. The third problem is worth investigating also because, due to the error discovered in a formula for regularisation parameter introduced in Sechidis et al. (2019), the behaviour of corresponding estimators is in need of re-examination. It turns out that theoretical advantages of SE-based regularisation are confirmed in the numerical experiments and the real data analysis.

The paper is organised as follows. In Sect. 2 we discuss the shrinkage estimators for the univariate and the bivariate case introduced in Hausser and Strimmer (2007) and Sechidis et al. (2019). In Sect. 3 we introduce a new proposal of regularisation parameter and discuss its estimation by crossvalidation method. Section 4 is devoted to discussion of some new properties of shrinkage estimators considered. Variable selection by information theoretic methods and regularized versions of selection criteria are discussed in Sect. 5. Section 6 contains results of numerical experiments on the behaviour of the shrinkage estimators of the distribution, the CMI as well as analysis of their use for Markov Blanket discovery. Section 7 concludes.

2 Shrinkage estimators: discussion of previous approaches

2.1 Shrinkage estimators of multinomial probabilities

We consider a probability distribution with $|\mathcal{X}| = m$ values characterized by a vector of probabilities $p = (p(x))_{x\in\mathcal{X}}$. The aim is to estimate p based on n values x_i of independent observations generated from this distribution summarized by a vector $N = (N(x))_{x\in\mathcal{X}}$ counting the number of respective values. Note that N has multinomial distribution $\mathrm{Mult}(n, p)$. We assume that $m > 1$ and $p(x) > 0$ for $x \in \mathcal{X}$. Let $\hat{p}^{(1)}$ and $\hat{p}^{(2)}$ be two estimators of p based on N. We assume, moreover, that $\hat{p}^{(2)}$ is unbiased i.e. $\mathbb{E}\,\hat{p}^{(2)} = p$. Usually, $\hat{p}^{(2)}$ is the maximum likelihood estimator of p which in this case is simply a fraction of observations assuming a corresponding value: $\hat{p}^{(2)}(x) = N(x)/n$. Note that when m is large compared to n some values of x

may not be represented among observations and this suggests that $\hat{p}^{(2)}$ may perform poorly in such cases. Estimator $\hat{p}^{(1)}$ usually involves less degrees of freedom than $\hat{p}^{(2)}$. In the extreme cases it can be a fixed distribution such as the uniform.

We define James–Stein shrinkage estimator of p pertaining to $\hat{p}^{(1)}$ and $\hat{p}^{(2)}$ as their convex combination

$$\hat{p}_\lambda = \lambda \hat{p}^{(1)} + (1-\lambda)\hat{p}^{(2)} \tag{1}$$

for appropriately chosen $\lambda \in [0, 1]$. Thus $\hat{p}_\lambda$ is the weighted average of these two estimators and the aim of the transformation is to balance their properties: a low variance and a high bias of $\hat{p}^{(1)}$ and lack of the bias but a higher variance of $\hat{p}^{(2)}$. The name 'shrinkage' used in this context is due to the observation that $\hat{p}^{(2)}$ is shrunk towards $\hat{p}^{(1)}$.

Traditionally considered measure of the goodness of fit of the above estimator is (global) mean SE (MSE)

$$MSE(\lambda) = \mathbb{E}\Big(\sum_x (p(x) - \hat{p}_\lambda(x))^2\Big) \tag{2}$$

and the corresponding choice of λ is defined by

$$\lambda_{MSE} = \arg\min_{\lambda\in[0,1]} MSE(\lambda). \tag{3}$$

Assume that $\mathbb{E}\,\hat{p}^{(2)}(x) = p(x)$. Then a simple calculation yields (see "Appendix") that provided $\hat{p}^{(1)} \not\equiv \hat{p}^{(2)}$

$$\begin{aligned}\lambda_{MSE} &= \frac{\sum_x \Big(\mathrm{Var}(\hat{p}^{(2)}(x)) - \mathrm{Cov}(\hat{p}^{(1)}(x), \hat{p}^{(2)}(x))\Big)}{\sum_x \mathbb{E}(\hat{p}^{(1)}(x) - \hat{p}^{(2)}(x))^2} && (4)\\ &= \frac{\sum_x \Big(\mathrm{Var}(\hat{p}^{(2)}(x)) - \mathrm{Cov}(\hat{p}^{(1)}(x), \hat{p}^{(2)}(x))\Big)}{\sum_x \Big[\mathbb{E}\left(\hat{p}^{(1)}(x)\right)^2 + \mathbb{E}\left(\hat{p}^{(2)}(x)\right)^2 - 2\,\mathbb{E}\,\hat{p}^{(1)}(x)\hat{p}^{(2)}(x)\Big]}. && (5)\end{aligned}$$

The formula above has been derived in Ledoit and Wolf (2003) for regularized covariance matrix estimation.

In the context of regularizing a vector of sample probabilities λ_{MSE} has been considered in Hausser and Strimmer (2007) for $\hat{p}^{(2)}(x)$ being ML estimator of $p(x)$, that is a fraction of samples equal to the corresponding value, and $\hat{p}^{(1)}$ equal to the uniform probability mass function $\hat{p}^{(1)}(x_i) = 1/m$, $i = 1, \ldots, m$. Note that we keep a 'hat' symbol over $p^{(1)}$ in $\hat{p}^{(1)}$ even though it does not depend on the data. In this case Formula (4) reduces to

$$\lambda^U_{MSE} = \frac{\sum_x \mathrm{Var}(\hat{p}^{(2)}(x))}{\sum_x \mathbb{E}(\hat{p}^{(1)}(x) - \hat{p}^{(2)}(x))^2} = \frac{1 - \sum_x p^2(x)}{n\sum_x \mathbb{E}(\hat{p}^{(1)}(x) - \hat{p}^{(2)}(x))^2} \tag{6}$$

as $\mathrm{Cov}(\hat{p}^{(1)}(x), \hat{p}^{(2)}(x)) = 0$. The superscript 'U' stands for the 'Uniform'.

We note in passing that λ^U_{MSE} can be interpreted in terms of variability index of distribution p. Namely, we have

Remark 1 It is easy to check from (6) that

$$\lambda^U_{MSE} = \frac{\sum_x \mathrm{Var}(\hat{p}^{(2)}(x))}{||U-p||^2 + \sum_x \mathrm{Var}(\hat{p}^{(2)}(x))}, \tag{7}$$

where $||U-p||^2$ is squared l^2 distance between $p(x)$ and the uniform distribution on $\{1,\ldots,m\}$. This can be written alternatively in terms of Gini indices of variability of discrete distributions $I_G(p) = 1 - \sum_x p^2(x)$, namely

$$\lambda^U_{MSE} = \frac{I_G(p)}{I_G(p) + n(I_G(U) - I_G(p))}, \tag{8}$$

where Gini index $I_G(U)$ for the uniform distribution attains the maximal value among Gini indices equal $1 - 1/m$. It also follows from (8) that $\lambda^U_{MSE} \sim 1/n$ provided distribution p is different from the uniform.

The most important alternative approach to regularisation is based on Bayesian paradigm and usually relies on imposing the Dirichlet prior on p. Namely, Bayesian regularisation of counts $N = (N(x))_{x\in X}$ is based on the assumption that prior p_{prior} for p is the Dirichlet distribution with hyperparameters $\alpha = (\alpha_x)_{x\in\mathcal{X}}$. In such a setting, the posterior is given in the following form

$$\hat{p}(x) = \frac{N(x)+\alpha_x}{n+A} = \frac{n}{n+A}\hat{p}^{(2)}(x) + \frac{A}{n+A}\frac{\alpha_x}{A},$$

where $A = \sum_x \alpha_x$. The right hand side of the equality above shows that the Bayesian estimator is a weighted average of the estimator $\hat{p}^{(2)}(x)$ and the expected value of the prior $\mathbb{E}\, p_{prior}(x) = A^{-1}\alpha_x$ and thus coincides with James–Stein-type shrinkage estimator. For a thorough comparison of the Bayesian and shrinkage approaches see Hausser and Strimmer (2007).

We define now plug-in estimator of λ^U_{MSE} in (6) as

$$\hat{\lambda}^U_{MSE} = \frac{1 - \sum_x \hat{p}^{(2)}(x)^2}{n\sum_x \hat{\mathbb{E}}(\hat{p}^{(2)}(x) - 1/m)^2}, \tag{9}$$

where

$$\hat{\mathbb{E}}(\hat{p}^{(2)}(x) - 1/m)^2 = \frac{n-1}{n}\hat{p}^{(2)}(x)^2 + \left(\frac{1}{n} - \frac{2}{m}\right)\hat{p}^{(2)}(x) + \frac{1}{m^2} \tag{10}$$

due to $\mathbb{E}(\hat{p}^{(2)}(x)^2) = p(x)((n-1)p(x)+1)/n$.

2.2 The bivariate vector case

In the following we focus on a case when vector $p = p(x, y)$ is a probability mass function of bivariate discrete random variable (X, Y) when X admits k values and Y admits l values. Thus $m = kl$ is the length of vector p. We assume that both k and l are larger then 1 and all m values are taken with non-zero probability. The approach is easily generalized to the case when X and Y are multivariate. Hausser-Strimmer's regularization method extends naturally to this case with $\hat{p}^{(1)}(x, y) = 1/m$. Recently (cf. Sechidis et al. 2019) $p_{\lambda}^{Ind}(x, y)$ has been introduced which equals (1) for $\hat{p}^{(2)}(x, y) = \hat{p}^{ML}(x, y)$ and

$$\hat{p}^{(1)}(x, y) = \hat{p}^{ML}(x) \times \hat{p}^{ML}(y) = \frac{n(x)}{n} \times \frac{n(y)}{n} \tag{11}$$

is ML estimator for $p(x, y)$ when X and Y are assumed independent. Intuitively, for this approach and cases close to independence the optimal λ should be close to 1 [compare (1)]. For such choice of $\hat{p}^{(1)}$ explicit form of λ_{MSE} denoted here by λ_{MSE}^{Ind} has been derived in Theorem 1 of Sechidis et al. (2019). However, it has been noted in Łazęcka and Mielniczuk (2020) that the formula for $\mathbb{E}\left(\hat{p}^{(1)}(x, y)\right)^2$ appearing in the denominator of (4) contains an error (see corrigendum in Sechidis 2020). Below we state for a future reference a corrected form of λ_{MSE}^{Ind}. In the "Appendix" we give a simple proof of the expression for $\mathbb{E}\left(\hat{p}^{(1)}(x, y)\right)^2$ which contained an error in the original publication.

Theorem 1 (corrected form of Theorem 1 in Sechidis et al. 2019) *If λ_{MSE}^{Ind} is defined as in (4) when $\hat{p}^{(2)}(x, y) = \hat{p}^{ML}(x, y)$ and $\hat{p}^{(1)}(x, y)$ equals (11), then the quantities appearing in (4) are equal*

$$\begin{aligned}
\mathrm{Var}(\hat{p}^{(2)}(x, y)) &= p(x, y)(1 - p(x, y))/n, \\
\mathbb{E}\left(\hat{p}^{(2)}(x, y)\right)^2 &= p(x, y)\left((n-1)p(x, y) + 1\right)/n, \\
\mathbb{E}\left(\hat{p}^{(1)}(x, y)\right)^2 &= \big((n-1)(n-2)(n-3)p^2(x)p^2(y) \\
&\quad +(n-1)(n-2)\left(p^2(x)p(y) + p(x)p^2(y)\right. \\
&\quad \left.+4p(x, y)p(x)p(y)\right) \\
&\quad +(n-1)\left(p(x)p(y) + 2p^2(x, y)\right. \\
&\quad \left.+2p(x, y)p(x) + 2p(x, y)p(y)\right) \\
&\quad +p(x, y)\big)/n^3, \mathrm{Cov}\left(\hat{p}^{(1)}(x, y), \hat{p}^{(2)}(x, y)\right) \\
&= p(x, y)\big((n-1)\left(p(x) + p(y)\right. \\
&\quad \left.-2p(x)p(y)\right) \\
&\quad +1 - p(x, y)\big)/n^2,
\end{aligned}$$

$$\mathbb{E}\left(\hat{p}^{(1)}(x,y)\hat{p}^{(2)}(x,y)\right) = p(x,y)\big((n-1)\big((n-2)p(x)p(y) + p(x) + p(y) + p(x,y)\big) + 1\big)/n^2.$$

Note that as λ_{MSE} depends on the unknown second moments of $\hat{p}^{(2)}$ and thus it needs to be estimated.

In the case of λ_{MSE}^{U} fractions are plugged in formulas for the moments of $\hat{p}^{(i)}(x,y)$, $i = 1, 2$ yielding

$$\hat{\lambda}_{MSE}^{U} = \lambda_{MSE} = \frac{\sum_{x,y} \widehat{\mathrm{Var}}(\hat{p}^{(2)}(x,y))}{\sum_{x,y} \widehat{\mathbb{E}}(\hat{p}^{(2)}(x,y) - 1/m)^2}, \tag{12}$$

where $\widehat{\mathrm{Var}}(\hat{p}^{(2)}(x,y))$ is obtained by plugging $\hat{p}^{(2)}$ into expressions pertaining to $\mathrm{Var}(\hat{p}^{(2)})$ (cf. Theorem 1) and analogously to (10)

$$\widehat{\mathbb{E}}\left(\hat{p}^{(2)}(x,y) - \frac{1}{m}\right)^2 = \frac{n-1}{n}[\hat{p}^{(2)}(x,y)]^2 + \left(\frac{1}{n} - \frac{2}{m}\right)\hat{p}^{(2)}(x,y) + \frac{1}{m^2} \tag{13}$$

as $\mathbb{E}\left(\hat{p}^{(2)}(x,y)\right)^2 = p(x,y)\left((n-1)p(x,y)+1\right)/n$. Moreover, $\hat{\lambda}_{MSE}^{Ind}$ is obtained by plugging in fractions into expressions for moments given in Theorem 1.

3 Shrinkage estimators: a new proposal based on stochastic accuracy

We consider now the (global) SE as an adopted measure of goodness of fit of an estimator, which is

$$SE(\lambda) = \sum_{x} (\hat{p}_{\lambda}(x) - p(x))^2 \tag{14}$$

and corresponding method of choosing λ, namely

$$\lambda_{SE} = \arg\min\nolimits_{\lambda\in[0,1]} SE(\lambda). \tag{15}$$

The rationale of minimising $SE(\lambda)$ is that our aim is to choose parameter λ of $\hat{p}_{\lambda}$ such that $\hat{p}_{\lambda}$ *is the best approximation of* $p(\cdot)$ *for the available data*. This is in contrast to choosing λ_{MSE} which is not data-dependent and which applied for many virtual samples drawn from $p(\cdot)$ will yield *the best average performance*.

The approach of choosing a regularisation parameter by minimising stochastic measure of accuracy is frequently applied. The primary example is density estimation (corresponding to the problem studied by us in the continuous case) for which a choice of smoothing (i.e. regularising) parameter is done based on minimisation of Integrated SE ISE as an alternative, to minimisation of mean ISE ($MISE$). In the seminal paper Stone (1984) considers the minimiser h_0 of ISE for the kernel estimator $\hat{f}_h$ of the underlying density f equal $ISE(h) = \int(\hat{f}_h(x) - f(x))^2\,dx$ as an object of primary interest and shows in his Theorem 1 that under suitable conditions a proposed estimator $\hat{h}_0$ of h_0 satisfies $\hat{h}_0/h_0 \to 1$ almost surely, thus $\hat{h}_0$ is asymptotically optimal. The behaviour of ISE as a stochastic accuracy measure was studied by Hall in a series of papers (e.g. Hall 1982, 1984). The subject was further pursued by Hall and Marron (1987) where they investigate interrelations between $ISE(\hat{h}_0)$ and $ISE(\hat{h})$, for $\hat{h}$ being e.g. cross-validatory window $\hat{h}_c$. The approach based on stochastic measures of accuracy has been since studied by many authors e.g. by Scott (2001) who proposed and investigated minimiser of $ISE(h)$ in parametric context. Other important contributions include Rice (1984), where minimisation of weighted ISE in regression is considered, Marron and Härdle (1986) where the corresponding problem is studied in a general framework of curve estimation and Sugiyama et al. (2012) among others.

Moreover, we stress that a popular crossvalidatory window $\hat{h}_c$ i.e. the minimiser of $CV(h) = \int \hat{f}_h^2(x)\,dx - 2n^{-1}\sum_{i=1}^n \hat{f}_{h,-i}(X_i)$, $((X_i)_{i=1}^n$ being the underlying sample and $f_{h,-i}(X_i)$ kernel estimator for the data with X_i omitted) *is based on minimisation of estimated $ISE(h)$ (and not $MISE$)*, as $CV(h)$ is estimator of $ISE(h) - \int f^2$ (see e.g. Scott 2001, Sect. 2 and Hall and Marron 1987). We view our approach as suitably tailored analogue of the approach above for probability mass function estimation. We stress that by minimising $SE(\lambda)$ we aim at approximation of the unknown density using available data, which is of primary interest.

We also observe that $m^{-1}SE(\lambda)$ can be written as

$$\mathbb{E}(p(U) - \hat{p}_\lambda(U))^2 | X_1, \ldots, X_n),$$

where U is uniformly distributed on the $\mathcal{X}$ and independent of $X_1, \ldots, X_n$, thus λ_{SE} yields the parameter of the best prediction in the squared sense for the randomly picked value from the range of X. This is reminiscent of the problem of assessing of the accuracy of the classifier which can be either done using conditional error which is a stochastic measure of accuracy and corresponds to the performance of the classifier built for the data at hand or unconditional accuracy which assesses the performance on many virtual copies of the data.

We note first that λ_{SE} defined in (15) can be easily calculated.

Lemma 1 *Shrinkage parameter λ_{SE} that minimizes squares error $SE(\lambda)$ given in (15) equals*

$$\lambda_{SE} = \lambda_{SE}(N) = \frac{\sum_x \left(\hat{p}^{(2)}(x) - p(x)\right)\left(\hat{p}^{(2)}(x) - \hat{p}^{(1)}(x)\right)}{\sum_x \left(\hat{p}^{(1)}(x) - \hat{p}^{(2)}(x)\right)^2}. \tag{16}$$

Proof Note that $SE(\lambda)$ is a quadratic function of λ. Namely,

$$SE(\lambda) = \sum_x (p(x) - \hat{p}_\lambda(x))^2 = \sum_x \Big[\lambda^2\left(\hat{p}^{(1)}(x) - \hat{p}^{(2)}(x)\right)^2 \\ -2\lambda\left(p(x) - \hat{p}^{(2)}(x)\right)\left(\hat{p}^{(1)}(x) - \hat{p}^{(2)}(x)\right) + \left(p(x) - \hat{p}^{(2)}(x)\right)^2\Big]. \tag{17}$$

Then, if $\hat{p}^{(1)}(x) \not\equiv \hat{p}^{(2)}(x)$, it is seen that the minimiser of the above function is indeed (16) by noting that $SE(\lambda)$ is a convex function and calculating its stationary point. □

We observe that for the derivation of λ_{SE} we do not assume that $\hat{p}^{(2)}(x)$ is an unbiased estimator of $p(x)$. In order to underline that λ_{SE} depends on multinomial variable N and thus it is a random quantity we will denote it by $\lambda_{SE}(N)$. We note that due to the definition of $\lambda_{SE}(N)$

$$\sum_x \left(\hat{p}_{\lambda_{SE}(N)}(x) - p(x)\right)^2 \leq \sum_x \left(\hat{p}_{\lambda_{MSE}}(x) - p(x)\right)^2$$

and thus taking expectations on both sides

$$MSE(\lambda_{SE}(N)) = \mathbb{E}_N \left(\sum_x \left(\hat{p}_{\lambda_{SE}(N)}(x) - p(x)\right)^2 \right) \leq MSE(\lambda_{MSE}), \qquad (18)$$

which is one of the main advantages of considering λ_{SE}. Namely, choosing the regularization parameter for every sample separately, we obtain the mean squared distance between the estimator $\hat{p}$ and the true vector p which is not larger than the minimal value of MSE for the *global* regularization parameter. We note that if $\hat{p}^{(1)}$ is the uniform distribution, then Formula (16) reduces to

$$\lambda_{SE}^{U} = \lambda_{SE}^{U}(N) = \frac{\sum_x \left(\hat{p}^{(2)}(x) - p(x)\right) \hat{p}^{(2)}(x)}{\sum_x \left(\hat{p}^{(1)}(x) - \hat{p}^{(2)}(x)\right)^2} \qquad (19)$$

as $\sum_x \left(p(x) - \hat{p}^{(2)}(x)\right) \hat{p}^{(1)}(x) = 0$, because $\hat{p}^{(1)}(x)$ does not depend on x. Note that $\lambda_{SE}^{U} = 1$ when $p(x) = \hat{p}^{(1)}(x)$ is uniform.

3.1 Crossvalidation estimators of λ_{SE}

In the case of λ_{SE} for both λ_{SE}^{U} and λ_{SE}^{Ind} we introduce cross-validation estimators. To this end we note that when ith observation x_i is omitted from the data set then letting $n(x) = \#\{i : x_i = x\}$, we have:

$$\hat{p}_{-i}^{(2)}(x) = \frac{n(x) - 1}{n - 1}$$

if $(x_i) = x$ and

$$\hat{p}_{-i}^{(2)}(x) = \frac{n(x)}{n - 1}$$

in the opposite case. Thus $\sum_x p(x)\hat{p}^{(2)}(x)$ is estimated by

$$\frac{1}{n}\sum_{i=1}^{n} \hat{p}_{-i}^{(2)}(x_i) = \sum_{x \in \mathcal{X}} \frac{n(x)}{n} \frac{n(x) - 1}{n - 1}.$$

As the mass assigned by $\hat{p}^{(2)}$ to x equals n^{-1}, the numerator of (19) is $n^{-1}\sum_{i=1}^{n}(\hat{p}^{(2)}(x_i)-p(x_i))$ and replacing $p(x_i)$ by its cross-validation estimator $\hat{p}^{(2)}_{-i}(x_i)$ we arrive at the following estimator for the numerator of (19)

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(\hat{p}^{(2)}(x_i)-\hat{p}^{(2)}_{-i}(x_i)\right)\\
&\quad=\sum_{x\in\mathcal{X}}\left(\frac{n^2(x)}{n^2}-\frac{n(x)}{n}\frac{n(x)-1}{n-1}\right)\\
&\quad=\frac{1}{n-1}\sum_{x\in\mathcal{X}}\frac{n\cdot n^2(x)-n^2(x)-n\cdot n^2(x)+n\cdot n(x)}{n^2}\\
&\quad=\frac{1}{n-1}\sum_{x\in\mathcal{X}}\frac{n(x)}{n}\left(1-\frac{n(x)}{n}\right),
\end{aligned}$$

from which the cross-validation estimator $\hat{\lambda}^{U}_{SE}$ is easily derived. The obtained form is intuitively clear as the expected value of the numerator of (19) is the global variance of $\hat{p}^{(2)}$. In the bivariate case we reason analogously for $\hat{\lambda}^{Ind}_{SE}$ with use of (16) and replacing $n(x)$ by $n(x,y)$. We thus have

$$\frac{1}{n}\sum_{i=1}^{n}\hat{p}^{(2)}_{-i}(x_i,y_i)=\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\frac{n(x,y)}{n}\frac{n(x,y)-1}{n-1}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{p}^{(2)}(x_i,y_i)-\hat{p}^{(2)}_{-i}(x_i,y_i)\right)=\frac{1}{n-1}\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\frac{n(x,y)}{n}\left(1-\frac{n(x,y)}{n}\right).$$

We then prove

Theorem 2 *Assume that $\hat{p}^{(1)}$ has the uniform distribution in the case of λ^{U}_{SE} and is defined in (11) for λ^{Ind}_{SE}. Cross-validation estimator of λ^{U}_{SE} and λ^{Ind}_{SE} are*

$$\hat{\lambda}^{U}_{SE}=\frac{\sum_{x\in\mathcal{X}}n(x)\,(n-n(x))}{n^2(n-1)\sum_{x\in\mathcal{X}}\left(1/m-\hat{p}^{(2)}(x)\right)^2}=\frac{\sum_{x\in\mathcal{X}}\hat{p}^{(2)}(x)(1-\hat{p}^{(2)}(x))}{(n-1)\sum_{x\in\mathcal{X}}\left(1/m-\hat{p}^{(2)}(x)\right)^2}\tag{20}$$

and

$$\hat{\lambda}^{Ind}_{SE}=\frac{\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\left(\begin{array}{r}\left(\hat{p}^{(2)}(x,y)\right)^2+\hat{p}^{(2)}(x,y)\left(\frac{(n(x)-1)(n(y)-1)}{(n-1)^2}-\frac{n(x,y)-1}{n-1}\right)\\-\hat{p}^{(2)}(x,y)\hat{p}^{(1)}(x,y)\end{array}\right)}{\sum_{x\in\mathcal{X},y\in\mathcal{Y}}\left(\hat{p}^{(1)}(x,y)-\hat{p}^{(2)}(x,y)\right)^2}.\tag{21}$$

Note that for $\hat{\lambda}_{SE}^{Ind}$ in the case of the numerator of (16) as an estimator of $\sum_{x,y} p(x,y)\big(\hat{p}^{(1)}(x,y) - \hat{p}^{(2)}(x,y)\big)$ we consider

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{p}_{-i}^{(1)}(x_i, y_i) - \hat{p}_{-i}^{(2)}(x_i, y_i)\right)$$
$$= \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \frac{n(x,y)}{n}\left(\frac{(n(x)-1)(n(y)-1)}{(n-1)^2} - \frac{n(x,y)-1}{n-1}\right).$$

It follows from a close scrutiny of the code for $\hat{\lambda}_{MSE}^{U}$ in Hausser and Strimmer (2014) that the authors use $\hat{\lambda}_{SE}^{U}$ in place of $\hat{\lambda}_{MSE}^{U}$ as they replaced the plug-in estimator of the expectation in the denominator of (12) by its unbiased sample counterpart (see Schäffer and Strimmer 2005). In our general approach we obtain $\hat{\lambda}_{SE}^{U}$ as an estimator of the parameter optimising SE accuracy measure.

4 Asymptotic behaviour of regularisation parameters $\hat{\lambda}$

We now state and prove theorems which show that regularisation parameters discussed above behave as expected, namely when $p(x)$ is not uniform $\hat{\lambda}_{MSE}^{U}$ and $\hat{\lambda}_{SE}^{U}$ tend to 0 in a specified sense when the sample size grows. This means that $\hat{p}_{\hat{\lambda}}$ becomes close to ML estimator $\hat{p}^{(2)}$ in this case. A similar statement holds for $\hat{\lambda}_{MSE}^{Ind}$ and $\hat{\lambda}_{SE}^{Ind}$. We also show that in the situations, where regularisation is especially beneficial (i.e. when the distribution of $p(x,y)$ is uniform or X and Y are independent), SE-based estimators of λ are greater than MSE-based estimators. Thus the regularisation weights assigned to the true model (in the first case) or a smaller subfamily containing the correct model (in the second case) are larger for SE-based than MSE-based estimators. In the appendix we state and prove analogous result for theoretical minimisers $\lambda_{MSE}^{U}, \lambda_{SE}^{U}$ and $\lambda_{MSE}^{Ind}, \lambda_{SE}^{Ind}$. When not specified otherwise, convergence is meant almost surely.

Theorem 3 *We have the following convergences provided $n \to \infty$:*

(i) *$n\hat{\lambda}_{MSE}^{U} \to c$ when $p(x) \not\equiv 1/m$ and otherwise $\hat{\lambda}_{MSE}^{U} \to \frac{m-1}{m-1+Q}$ in distribution, where $c > 0$ is defined in (22) and Q has χ_{m-1}^2 distribution;*

(i) *$n\hat{\lambda}_{SE}^{U} \to c$ when $p(x) \not\equiv 1/m$ and otherwise $\hat{\lambda}_{SE}^{U} \to \frac{m-1}{Q}$, where Q has χ_{m-1}^2 distribution.*

Moreover, we have that $\hat{\lambda}_{SE}^{U} \geq \hat{\lambda}_{MSE}^{U}$.

Proof The first part of (i) follows from (9) and (10) as they imply that n times the numerator of $\hat{\lambda}_{MSE}^{U}$ tends to $\sum_x p(x)(1-p(x))$ whereas the denominator tends to $\sum_x (p(x)-1/m)^2 > 0$ [cf. equality (10)]. Thus we have that the conclusion of the first part of (i) holds with

$$c = \frac{\sum_x p(x)(1-p(x))}{\sum_x (p(x)-1/m)^2} > 0. \tag{22}$$

In the case $p(x) \equiv 1/m$, we have that n times the numerator tends to $1 - \sum_x p^2(x) = 1 - 1/m$ as before and n times the denominator equals [cf. equality (10)]

$$n \sum_x \hat{\mathbb{E}} \left(\hat{p}^{(2)}(x) - \frac{1}{m} \right)^2 \\ = \sum_x \hat{p}^{(2)}(x) \left(1 - \hat{p}^{(2)}(x) \right) + n \sum_x \left(\hat{p}^{(2)}(x) - \frac{1}{m} \right)^2 . \tag{23}$$

The first sum tends to $\sum_x p(x)(1 - p(x))$ and the second can be written as

$$n \sum_x \left(\hat{p}^{(2)}(x) - \frac{1}{m} \right)^2 = \frac{1}{m} \sum_x n \frac{\left(\hat{p}^{(2)}(x) - 1/m \right)^2}{1/m} =: \frac{X^2}{m}, \tag{24}$$

where $X^2 = mn(\hat{p}^{(2)}(\cdot) - p(\cdot))^T (\hat{p}^{(2)}(\cdot) - p(\cdot))$ and $\hat{p}^{(2)}(\cdot)$ denotes the vector of ML estimators for $p(\cdot)$. Thus under the assumption that $p(x) \equiv 1/m$, X^2 tends to χ^2 distribution with $m - 1$ degrees of freedom. This follows by noticing that X^2 coincides with n times chi square statistic for testing goodness of fit with the uniform distribution and well known result concerning asymptotic distribution of chi square statistic (cf. e.g. Bartoszyński and Niewiadomska-Bugaj 1996, Theorem 17.2.1).

The first part of (i) can easily obtained from (20) as

$$n \hat{\lambda}_{SE}^U = \frac{n}{n-1} \frac{\sum_x \hat{p}^{(2)}(x) \left(1 - \hat{p}^{(2)}(x) \right)}{\sum_x \left(\hat{p}^{(2)}(x) - 1/m \right)^2} \to c,$$

where c is the constant defined in (22). The second part follows from (24) and the form of $\hat{\lambda}_{SE}^U$

$$\hat{\lambda}_{SE}^U = \frac{n}{n-1} \frac{\sum_x \hat{p}^{(2)}(x) \left(1 - \hat{p}^{(2)}(x) \right)}{n \sum_x \left(\hat{p}^{(2)}(x) - 1/m \right)^2} \to \frac{1 - 1/m}{Q/m} = \frac{m-1}{Q}.$$

The inequality $\hat{\lambda}_{SE}^U \geq \hat{\lambda}_{MSE}^U$ reduces to the comparison of denominators of $\hat{\lambda}_{SE}^U$ and $\hat{\lambda}_{MSE}^U$ as both nominators are equal $\sum_x \hat{p}^{(2)}(x) \left(1 - \hat{p}^{(2)}(x) \right)$ and $n/(n-1) > 1$. We immediately have

$$n \sum_x \left(\hat{p}^{(2)}(x) - 1/m \right)^2 < \sum_x \hat{p}^{(2)}(x) \left(1 - \hat{p}^{(2)}(x) \right) + n \sum_x \left(\hat{p}^{(2)}(x) - \frac{1}{m} \right)^2 ,$$

which completes the proof in view of (20) and (23). □

Important corollary of the result is that when $p(x) \not\equiv 1/m$ the limiting distribution of $\hat{p}_{\hat{\lambda}}$ coincides with that of $\hat{p}^{(2)}$.

Corollary 1 (i) *Let $\hat{\lambda}$ denote either $\hat{\lambda}^U_{MSE}$ or $\hat{\lambda}^U_{SE}$ and $\hat{p}_{\hat{\lambda}}(\cdot)$ vector of corresponding regularised estimators. Then we have when $p(x) \not\equiv 1/m$ that*

$$\sqrt{n}(\hat{p}_{\hat{\lambda}}(\cdot) - p(\cdot)) \to N(0, \Sigma),$$

where $\Sigma = n\Sigma_{\hat{p}^{(2)}}$.
(ii) *When $p(x) \equiv 1/m$ then $\sqrt{n}(\hat{p}_{\hat{\lambda}}(\cdot) - p(\cdot))$ with $\lambda = \hat{\lambda}^U_{MSE}$ is asymptotically equivalent to*

$$\frac{mn[(\hat{p}^{(2)}(\cdot) - p(\cdot))^T(\hat{p}^{(2)}(\cdot) - p(\cdot))]\sqrt{n}(\hat{p}^{(2)}(\cdot) - p(\cdot))}{(m-1) + mn[(\hat{p}^{(2)}(\cdot) - p(\cdot))^T(\hat{p}^{(2)}(\cdot) - p(\cdot))]}$$

and then $\sqrt{n}(\hat{p}_{\hat{\lambda}}(\cdot) - p(\cdot))$ with $\lambda = \hat{\lambda}^U_{SE}$ is equivalent to

$$\frac{mn[(\hat{p}^{(2)}(\cdot) - p(\cdot))^T(\hat{p}^{(2)}(\cdot) - p(\cdot))] - (m-1)}{mn[(\hat{p}^{(2)}(\cdot) - p(\cdot))^T(\hat{p}^{(2)}(\cdot) - p(\cdot))}\sqrt{n}(\hat{p}^{(2)}(\cdot) - p(\cdot)).$$

Part (i) easily follows from decomposition

$$\sqrt{n}(\hat{p}_{\hat{\lambda}}(\cdot) - p(\cdot)) = \sqrt{n}\hat{\lambda}(\hat{p}^{(1)}(\cdot) - p(\cdot)) + \sqrt{n}(1 - \hat{\lambda})(\hat{p}^{(2)}(\cdot) - p(\cdot)),$$

the fact that $\sqrt{n}\hat{\lambda} \to 0$ proved above, Slutzky lemma and asymptotic distribution of ML estimator $\hat{p}^{(2)}$. Part (ii) follows by simple calculation noticing that in this case $\hat{\lambda}(p^{(1)}(\cdot) - p(\cdot)) \equiv 0$. Note, moreover, that it follows that $\sqrt{n}(\hat{p}_{\hat{\lambda}}(\cdot) - p(\cdot))$ with $\lambda = \hat{\lambda}^U_{MSE}$ is equivalent to $Z \times X^2/((m-1) + X^2)$ where X^2 is asymptotically χ^2_{m-1} and Z is $N(0, \Sigma)$. Analogous representation for $\sqrt{n}(\hat{p}_{\hat{\lambda}}(\cdot) - p(\cdot))$ with $\lambda = \hat{\lambda}^U_{MSE}$ is $Z \times (X^2 - (m-1))/X^2$.

Remark 2 Note that if $p(x) \equiv 1/m$ then $\hat{\lambda}^U_{SE}$ tends in distribution to scaled inverse chi-squared distribution $T = (m-1)/Q$ with number of degrees of freedom $m - 1$ and scaling parameter 1. In particular $\mathbb{E}\,T = \frac{m-1}{m-3}$. This easily follows from the form of $\hat{\lambda}^U_{SE} = \frac{m-1}{Q}$, where Q tends to χ^2_{m-1}, Slutzky lemma and definition of the scaled inverse chi squared distribution.

We now state the analogous result for regularisation parameters $\hat{\lambda}^{Ind}_{MSE}$ and $\hat{\lambda}^{Ind}_{SE}$.

Theorem 4 *We have the following convergences provided $n \to \infty$:*

(i) *$n\hat{\lambda}^{Ind}_{MSE} \to c$ when $p(x, y) \not\equiv p(x)p(y)$, c is defined in (27), and otherwise $\hat{\lambda}^{Ind}_{MSE}$ tends in distribution to $\frac{a}{a+\tilde{Q}}$, where $a = \sum_{x,y} p(x)(1 - p(x))p(y)(1 - p(y))$, $\tilde{Q}$ is distributed as $Z'AZ$, $Z \sim \mathcal{N}(0, \Sigma)$, Σ is defined in the Corollary 1 and A is a matrix of coefficients defined in (28).*
(ii) *$n\hat{\lambda}^{Ind}_{SE} \to c$ when $p(x, y) \not\equiv p(x)p(y)$ and $\hat{\lambda}^{Ind}_{SE} \to \frac{a}{\tilde{Q}}$ otherwise, where c, a and $\tilde{Q}$ are defined in* (i).

Proof First, we give both estimators in an expanded form. Below we use the notation $\hat{p}^{(1)}(x,y)=\hat{p}^{(1)}(x)\hat{p}^{(1)}(y)$. We have (cf. Theorem 2)

$$\hat{\lambda}^{Ind}_{MSE}=\frac{\hat{N}^{Ind}_{MSE}}{\hat{D}^{Ind}_{MSE}}, \tag{25}$$

where

$$\begin{aligned}
\hat{N}^{Ind}_{MSE} &= \sum_{x,y}\hat{p}^{(2)}(x,y)\left(1-\hat{p}^{(2)}(x,y)-\hat{p}^{(1)}(x)-\hat{p}^{(1)}(y)+2\hat{p}^{(1)}(x,y)\right)/n\\
&\quad +O(1/n^2),\\
\hat{D}^{Ind}_{MSE} &= \sum_{x,y}\left(\hat{p}^{(2)}(x,y)-\hat{p}^{(1)}(x,y)\right)^2+\Big(\hat{p}^{(2)}(x,y)-6(\hat{p}^{(1)}(x,y))^2\\
&\quad +\hat{p}^{(1)}(x,y)(\hat{p}^{(1)}(x)+\hat{p}^{(1)}(y))+10\hat{p}^{(1)}(x,y)\hat{p}^{(1)}(x)\hat{p}^{(1)}(y)\\
&\quad -2\hat{p}^{(2)}(x,y)(\hat{p}^{(1)}(x)+\hat{p}^{(1)}(y))-3(\hat{p}^{(2)}(x,y))^2\Big)/n+O(1/n^2)
\end{aligned}$$

and

$$\hat{\lambda}^{Ind}_{SE}=\frac{\hat{N}^{Ind}_{SE}}{\hat{D}^{Ind}_{SE}}, \tag{26}$$

where

$$\begin{aligned}
\hat{N}^{Ind}_{SE} &= \frac{n}{(n-1)^2}\sum_{x,y}\hat{p}^{(2)}(x,y)\left(1-\hat{p}^{(2)}(x,y)-\hat{p}^{(1)}(x)-\hat{p}^{(1)}(y)+2\hat{p}^{(1)}(x,y)\right)\\
&\quad +\frac{1}{(n-1)^2}\sum_{x,y}\hat{p}^{(2)}(x,y)\left(\hat{p}^{(2)}(x,y)-\hat{p}^{(1)}(x,y)\right),\\
\hat{D}^{Ind}_{SE} &= \sum_{x,y}\left(\hat{p}^{(2)}(x,y)-\hat{p}^{(1)}(x,y)\right)^2.
\end{aligned}$$

This easily follows from a form of sample analogues of the expressions given in Theorem 2 and recalling that all estimators of probabilities are bounded by 1.
The first part of both (i) and (i) is a conclusion from above equations, as n times both nominators tend to $\sum_{x,y}p(x,y)\,(1-p(x,y)-p(x)-p(y)+2p(x)p(y))$ and denominators tends to $\sum_{x,y}(p(x,y)-p(x)p(y))^2>0$. Thus

$$c=\frac{\sum_{x,y}p(x,y)\,(1-p(x,y)-p(x)-p(y)+2p(x)p(y))}{\sum_{x,y}(p(x,y)-p(x)p(y))^2}. \tag{27}$$

In the second part of (i) ($p(x,y) \equiv p(x)p(y)$)

$$n\hat{N}_{MSE}^{Ind} \to \sum_{x,y} p(x)p(y)\,(1 - p(x) - p(y) + p(x)p(y))$$
$$= \sum_{x,y} p(x)p(y)(1-p(x))(1-p(y)) =: a$$

and

$$n\hat{D}_{MSE}^{Ind} \to \tilde{Q} + \sum_{x,y} \Big(p(x)p(y) - 6p^2(x)p^2(y) + p(x)p(y)(p(x)+p(y))$$
$$+ 10p^2(x)p^2(y) - 2p(x)p(y)(p(x)+p(y)) - 3p^2(x)p^2(y) \Big)$$
$$= \tilde{Q} + \sum_{x,y} p(x)p(y)\Big(1 - p(x) - p(y) + p(x)p(y)\Big)$$
$$= \tilde{Q} + \sum_{x,y} p(x)p(y)(1-p(x))(1-p(y)),$$

where $n\sum_{x,y}\left(\hat{p}^{(2)}(x,y) - \hat{p}^{(1)}(x,y)\right)^2 \to \tilde{Q}$. This follows from the delta method (cf. e.g. Agresti 2013) applied to a function

$$f(\mathbf{p}) = \sum_{x,y}(p(x,y) - p(x)p(y))^2,$$

where $\mathbf{p} = (p(x_1, y_1), \ldots p(x_k, y_l))$. Then

$$A = \left(\frac{1}{2} \frac{\partial^2 f(\mathbf{p})}{\partial p(x,y) \partial p(x',y')} \right)_{(x,y),(x',y')}. \tag{28}$$

The second part of (i) is obvious. □

Remark 3 Note that the analogous statement to Corollary 1 (i) holds for regularised estimators based on λ_{MSE}^{Ind} and λ_{SE}^{Ind}. Moreover, although inequality $\lambda_{SE}^{Ind} \geq \lambda_{MSE}^{Ind}$ does not necessarily hold, it holds for their distributional limits as $a/\tilde{Q} \geq a/(a+\tilde{Q})$. We also note that it follows from the last two theorems that in the case when $p(x) \equiv 1/m$ or $p(x,y) \equiv p(x)p(y)$, $\hat{\lambda}_{SE}^{U}$ and $\hat{\lambda}_{SE}^{Ind}$ respectively may exceed 1 with non-zero probability. In such cases both estimators are truncated at 1 in Sect. 6.

5 Variable selection using information theoretic approach

We refer to Cover and Thomas (2006) for basic information theoretic concepts such as the mutual information MI and its conditional counterpart CMI used below.
Let Y be the discrete target variable and $X = (X_1, \ldots, X_p)$ the vector of discrete potential predictors. We will denote by X_S for $S \subset \{1, \ldots, p\}$ a subvector of X with

indices contained in S. Assume that at a certain stage of a variable selection procedure, the subset X_S has been already chosen as a vector of useful predictors and we would like to pick a next variable among remaining ones which contributes the most to understanding of Y. One of the most intuitive approaches would be to add a variable whose inclusion gives the most significant improvement of the mutual information, i.e. we find

$$\arg\max_{j\in S^c}\left[I(X_{S\cup\{j\}},Y)-I(X_S,Y)\right]=\arg\max_{j\in S^c} I(X_j,Y|X_S), \tag{29}$$

where CMI $I(X_j, Y|X_S)$ is defined as

$$I(X_j,Y|X_S)=\sum_{x,y,x_S} p(x,y,x_S)\log\left(\frac{p(x,y|x_S)}{p(x|x_S)p(y|x_S)}\right) \tag{30}$$

and $p(x, y, x_s) = P(X = x, Y = y, X_S = x_S)$, $p(x, y|x_S) = P(X = x, Y = y|X_S = x_S)$. The main problem with applying this approach is that estimation of $I(X_j, Y|X_S)$ becomes more erratic when S grows. Indeed, in this case we have to estimate conditional probability $p(x, y|x_S)$ on each strata $X_S = x_S$ for a fixed number of observations while the number of such strata grows exponentially when new predictors are added to X_S. In practise lower order criteria which approximate CMI in some sense are used in its place in (29). In particular, expanding CMI using Möbius expansion (cf. e.g. Meyer et al. 2008) and deleting all expansion terms of order higher than 2 or 3 is frequently used (see e.g. Brown et al. 2012). In this way CIFE (conditional infomax feature extraction, Lin and Tang 2006) criterion and JMI, Yang and Moody 1999) criterion of order 2 defined below are obtained. It has been observed in Brown et al. (2012) that such criteria can be written in a general form

$$J^{\beta,\gamma}(X_j,Y|X_S)=I(X_j,Y)-\beta\sum_{i\in S}I(X_i,X_j)+\gamma\sum_{i\in S}I(X_i,X_j|Y). \tag{31}$$

The respective terms above are known as relevancy, redundancy and complementarity of X_k. In particular CIFE criterion equals

$$CIFE(X_j)=I(X_j,Y)+\sum_{i\in S}[I(X_i,X_j|Y)-I(X_i,X_j)] \tag{32}$$

and corresponds to $(\beta, \gamma) = (1, 1)$. Also, for JMI criterion

$$JMI(X_j)=I(X_j,Y)+\frac{1}{|S|}\sum_{i\in S}[I(X_i,X_j|Y)-I(X_i,X_j)]=\frac{1}{|S|}\sum_{i\in S}I(X_j,Y|X_i)$$

$(\beta, \gamma) = (1/|S|, 1/|S|)$. JMI was also proved to be an approximation of CMI under certain dependence assumptions (cf. Vergara and Estevez 2014). mRMR criterion (cf. Peng et al. 2005) corresponds to $(\beta, \gamma) = (|S|^{-1}, 0)$ whereas more general MIFS criterion (Battiti 1994) pertains to pair $(\beta, 0)$. Obviously, the simplest univariate filter criterion known as MIM (mutual information maximisation Lewis 1992) corresponds to (0, 0) pair. We consider the sample version of JMI and CIFE which are the special cases of a sample versions of of $J^{\beta,\gamma}(X_j)$ defined as

$$\hat{J}^{\beta,\gamma}(X_j) = \hat{I}(X_j, Y) - \beta \sum_{i \in S} \hat{I}(X_i, X_j) + \gamma \sum_{i \in S} \hat{I}(X_i, X_j|Y), \tag{33}$$

which is obtained by plugging in fractions as an estimators of probabilities as well as its regularised version using shrinkage estimators. We also consider the following generalisation of JMI defined above which includes the third order terms. Namely, we note that

$$|S| JMI(X_k) = \sum_{i \in S} I(X_i, X_k; Y) - \sum_{i \in S} I(X_i, Y),$$

where the second does not depend on X_k and define

$$JMI3(X_k) = \sum_{\{i,j\} \subset S} I(X_i, X_j, X_k; Y) \tag{34}$$

where summation is over all subsets $\{i, j\} \subset S$.
As a reference we consider also CMIM (conditional mutual information maximization) method introduced in Fleuret (2004) which uses minimax criterion to choose a candidate and for which criterion function is defined as

$$CMIM(X_k) = \min_{j \in S} I(Y; X_k|X_j). \tag{35}$$

We note that for k-best subset selection of predictors complexity of JMI and CIFE scales as $O(k^2 p)$ whereas for JMI3 it scales as $O(k^3 p)$ (see Vinh et al. 2016).
We follow the approach introduced in Sechidis et al. (2019) and regularise ML estimator of $p(x_k, y, x_S)$ appearing in a plug-in version of $I(X_k, Y|X_j)$ using the method described in the previous section. More specifically, it relies on a low-dimensional approximation $p^{(1)}$ which is either the uniform distribution $p^U(x_j, y, x_S)$ or a distribution corresponding to conditionally independent X_j and Y given X_S. Thus for each criterion functions we consider four methods of regularising probability distributions: `unif` (uniform $p^{(1)}$, λ chosen by MSE minimisation), `unif.se` (uniform $p^{(1)}$, λ chosen by SE minimisation), analogously defined `indep` and `indep.se`, and compare it with the benchmark `ML` (maximum likelihood estimator with no regularisation).

6 Simulation study

The aim of the simulation study is to compare the performance of the procedures discussed above for estimating the probability mass function, the mutual information and the CMI in terms of MSE. We also investigated if the proposed procedures improve performance of variable selection criteria. The code for all implemented procedures as well as the results of additional numerical experiments are available in online supplement.[1]

6.1 Artificial data

We performed simulations for three Archimedean copulas (Clayton, Gumbel and Frank) and normal copula for a grid of parameters which in each case correspond to the whole range of modelled dependence. We refer to Nelsen (2006) for introduction to copulas. The normal copula is parametrized by dependence parameter $0 \leq \rho < 1$ which results in AR(1) dispersion structure $\Sigma = (\sigma_{ij})$, where $\sigma_{ij} = \rho^{|i-j|}$. The copulas were discretized into k bins in each dimension. For every dimension probability mass function of a marginal distribution is a mixture of the uniform distribution and binomial distribution $Bin(k-1, 1/2)$ with parameter $k-1$ and probability $p = 1/2$. Thus the marginal distribution has the following form:

$$P(X = i) = \alpha \binom{k-1}{i-1} \left(\frac{1}{2}\right)^{k-1} + (1-\alpha)\frac{1}{k} \quad \text{for} \;\; i = 1, 2, \ldots, k, \tag{36}$$

where $\alpha \in [0, 1]$. Note that α controls deviation from the uniform distribution: for small α the marginal distributions are close to the uniform. Parameter α has been introduced to investigate influence of the marginal distributions on the behaviour of `unif` and `indep` versions of regularisation parameters. Through discretization according to (36) for any dimension we then obtained discrete probabilities $p(x, y)$ (for two-dimensional copulas) and $p(x, y, z)$ (for three-dimensional) we sampled from. Every experiment was repeated $N = 200$ times.

The second model we considered was based on the Dirichlet distribution supported on probability vectors. For a chosen value of β we sampled from the Dirichlet distribution $D(\beta_1, \ldots, \beta_K)$ with all $\beta_i = \beta$, K-dimensional vectors of probabilities $p(x, y)$ or $p(x, y, z)$, where $K = k \times k$ (bivariate case) or $K = k \times k \times k$ (trivariate case). For equal values of parameters and $\beta > 1$, the samples for which all values are close to each other are more likely to be generated, thus the probability distributions close to the uniform are most frequently chosen. For each β we sampled $N = 200$ probability vectors p and then for a chosen mixing parameter α, in the case of $p^{(1)}$ corresponding to independence we considered in the two-dimensional case:

$$\tilde{p}(x, y) = \alpha p(x, y) + (1-\alpha) p(x) p(y)$$

and in three-dimensional case:

$$\tilde{p}(x, y, z) = \alpha p(x, y, z) + (1-\alpha) p(x, z) p(y).$$

For $p^{(1)}$ being the uniform distribution the second term equals $(1-\alpha)/K$ in both cases.

[1] github: *lazeckam/SE_shrinkEstimator*.

As a measure of performance of the estimators under study we considered estimated MSE, which in the case of regularised probability estimators equals $\sum_x \mathbb{E}(\hat{p}_{\hat{\lambda}}(x) - p(x))^2$ and is defined analogously for other parameters.

6.1.1 Results: estimation of p, $I(X, Y)$ and $I(X, Y|Z)$

Figures 1, 2 and 3 show the results for the copula examples. Figure 1 exhibits estimated MSE for probability mass function $p(x, y)$ for $n = 100$, $k = 10$ and two-dimensional copulas. Rows correspond to different values of α in (36) ($\alpha = 0, 0.5$ and 1). Shrinkage estimators outperform `ML` in all presented cases and the SE-based procedures `unif.se` and `indep.se` perform better than `unif` and `indep`. The largest differences between the ML and the shrinkage estimators occur for parameters that correspond to independence between variables, and they decrease when the dependence increases. We also observe that for the uniform marginal distributions ($\alpha = 0$) `unif` procedures outperform `indep` ones with `unif.se` outperforming `unif` (for $\alpha > 0$ differences between `unif.se` and `unif` are smaller). However, the advantage of `indep` methods is evident when the marginal distributions are binomial ($\alpha = 1$) with `indep.se` being the winner for all copulas in this case. Figure 2 also presents estimated MSE for the probability mass function p in case of three-dimensional copulas ($n = 125$, $k = 5$). Note that in both cases the parameters have been chosen so that the average number of observations per cell is 1. In this case estimator $\hat{p}^{(1)}(x, y)$ for regularisation parameters `ind` and `ind.se` is defined as ML estimator calculated under independence of (X, Z) and Y. The conclusions concerning MSE for p are consistent with the previous example, with differences between estimators being even more pronounced. Moreover, in that scenario we can also evaluate MSE for the CMI $I(X, Y|Z)$, which is shown in Fig. 3. The values of ML estimator are truncated from above at 0.03 and the vertical line corresponds to independence. The results for $I(X, Y|Z)$ indicate that replacing the ML estimator with the shrinkage procedures, in particular with the crossvalidation estimators based on SE minimisation, can enhance the performance of estimation in terms of MSE. However, the positive effect of SE minimisation methods is not always as pronounced for CMIs as for estimation of probability mass function, and there are some cases, as e.g. for the normal copula for $\alpha > 0$ and $\rho > 0.5$ when performance of `indep` is better than `indep.se`. This is due to to the fact that parameter λ is selected to minimise error for the probabilities and not for the CMI. Nevertheless, in many cases the superiority of shrinkage estimators over ML and shrinkage estimators based on SE minimisation over the remaining ones is evident. For larger n with fixed k the differences between estimators become less obvious. More results can be found in the online supplement and they include:

- plots of MSE for mutual information for the same n and k as above both for two and three-dimensional copulas,
- plots of MSE for p for two-dimensional copulas for $n = 100$ and $k = 5, 15, 20$; $k = 10$ and $n = 100, 500$; $n = 400$ and $k = 20$,
- plots of MSE for p and CMI for three-dimensional copulas for $n = 125$ and $k = 10, 15$; $n = 5$ and $k = 250$.

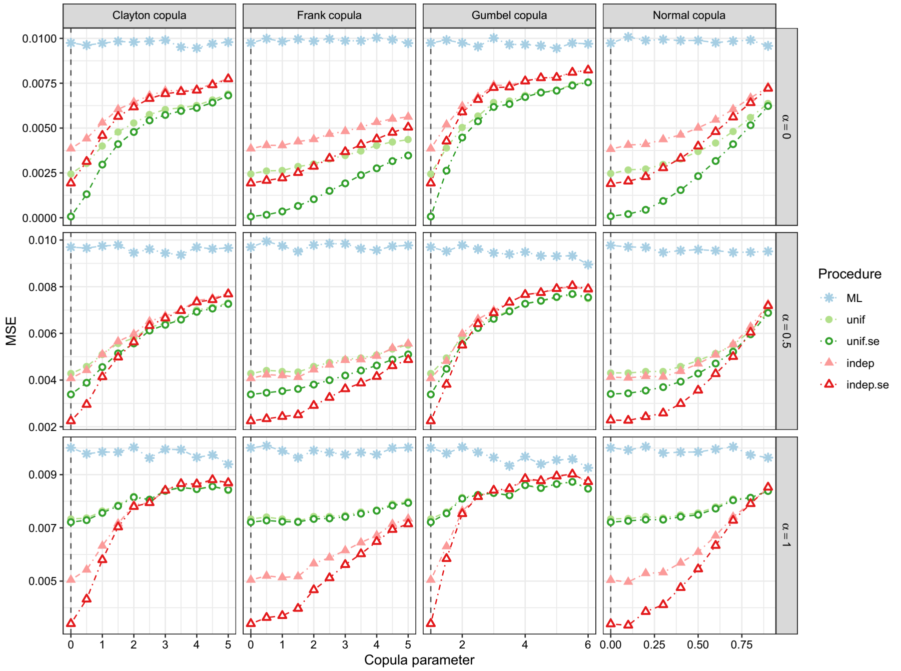


Fig. 1 MSE of estimated probabilities $p(x, y)$ for the copula models. X and Y variables have $k = 10$ levels each, their distribution being discretised copulas, the number of observation is $n = 100$. Copula parameters indicating independence are marked with vertical dashed lines

The conclusions for the results presented in the supplement are consistent with those stated above.

In Figs. 4 and 5 the results for the second model based on the Dirichlet distribution are presented. Figure 4 shows that all shrinkage estimators outperform ML with SE-based estimators being better that MSE-based estimators in terms of MSE, both in two-dimensional (Fig. 4a) and three-dimensional (Fig. 4b) scenarios. In almost all the cases considered unif.se works best, as the true distribution for the symmetric Dirichlet distribution (with all parameters being equal to β) with $\beta > 1$ is close to the uniform. Figure 5 shows the behaviour of estimators for mutual information (left panel, two-dimensional case) and CMI (right panel, three-dimensional case). Here all methods using regularisation perform similarly with indep.se working best for $\alpha = 0$ and $\alpha = 0.5$ and indep being superior for smaller values of parameter β of Dirichlet distribution and $\alpha = 1$.

6.2 Bayesian networks

In this section we consider benchmark Bayesian networks to measure the performance of shrinkage estimators when incorporated into popular second order variable selection criteria: JMI, CIFE and CMIM defined above. In our study we also included a third-order criterion: JMI3 defined in (34) and proposed in Sechidis et al. (2019). The

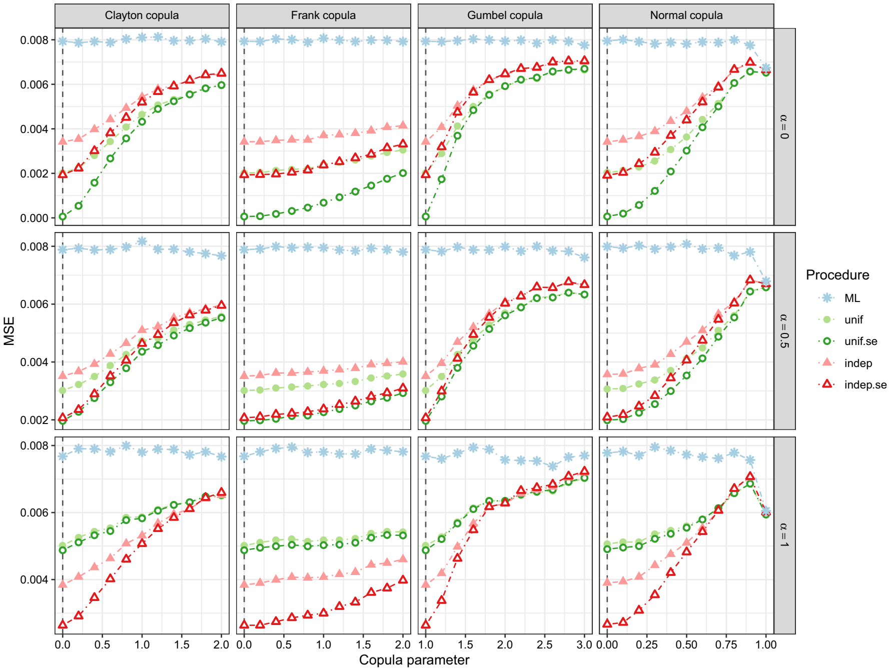


Fig. 2 MSE calculated for estimated probabilities $p(x, z, y)$ for the copula models. X, Y and Z have $k = 5$ levels, the number of observation is $n = 125$. Copula parameters indicating independence are marked with vertical dashed lines. In some cases results for `unif` and `unif.se` almost coincide

objective is to determine, using these methods, the Markov Blanket $MB(Y)$ for the given node Y, i.e. set of such variables that Y and remaining nodes are conditionally independent given $MB(Y)$ (see e.g. Brown et al. 2012). The main advantage of sampling from known networks is that for each node Y its Markov Blanket is known, therefore we can compare sets of chosen variables with the corresponding true Markov Blankets. In our study we include all of the small, medium and large networks available at the `bnlearn` package repository (cf. Scutari 2010) which has at least one node that satisfies the following condition: it has at least one child, one parent and one spouse (the same condition was imposed in Sechidis et al. 2019). We also include one network that is labelled as very large (`pathfinder`).

The simulations scenario was as follows: for each dataset we selected all nodes that satisfied the above mentioned condition and using each of them as the target variable and the remaining nodes as the explanatory variables, we applied the variable selection criteria combined with estimation procedures considered in previous sections. For each criterion we set the common number of chosen variables that was equal to the actual size of the Markov Blanket. The performance of each criterion and each estimation procedure was assessed using true positive rate (TPR, Recall). Then TPR was averaged over target nodes and simulations to obtain the result for each dataset. Note that in this case, due to the way the number of the chosen variables is determined, Precision

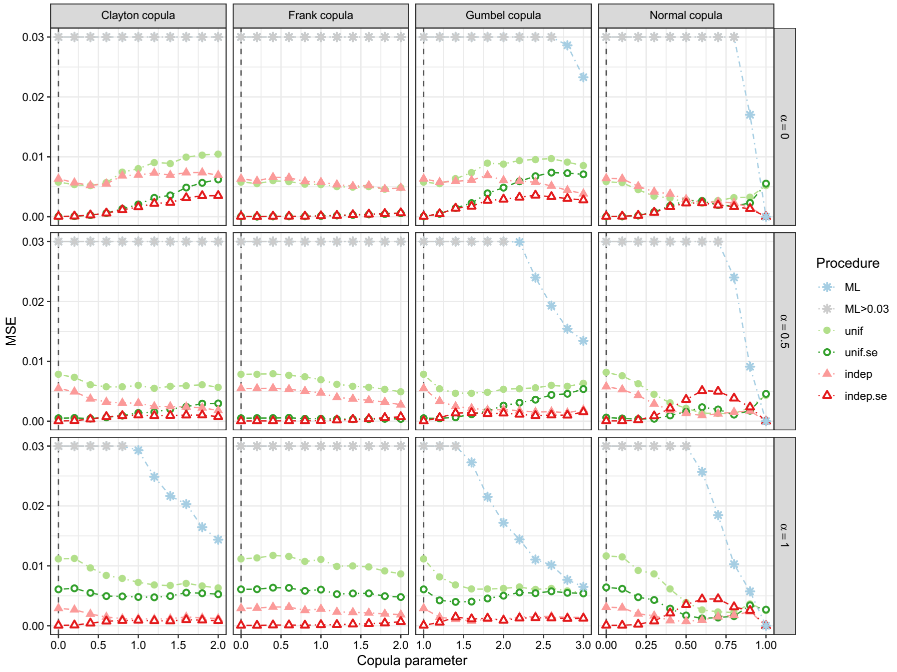
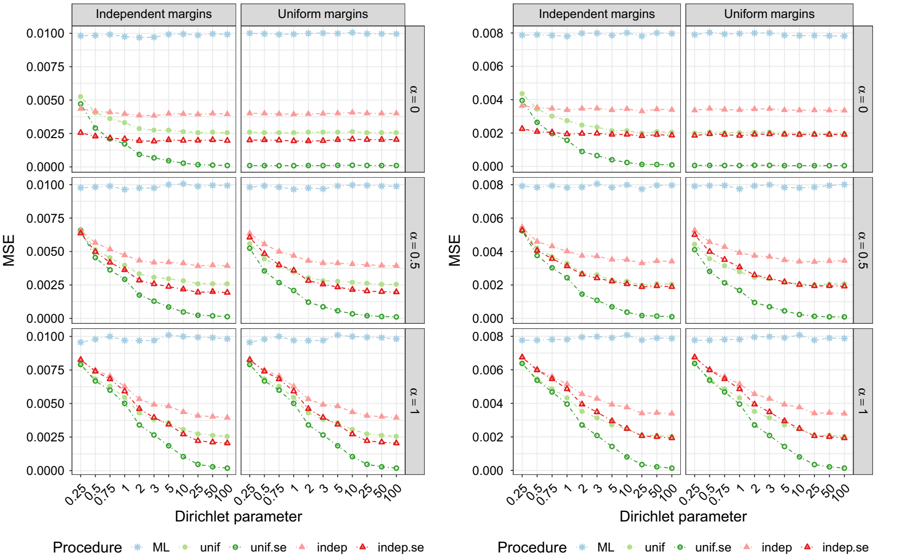


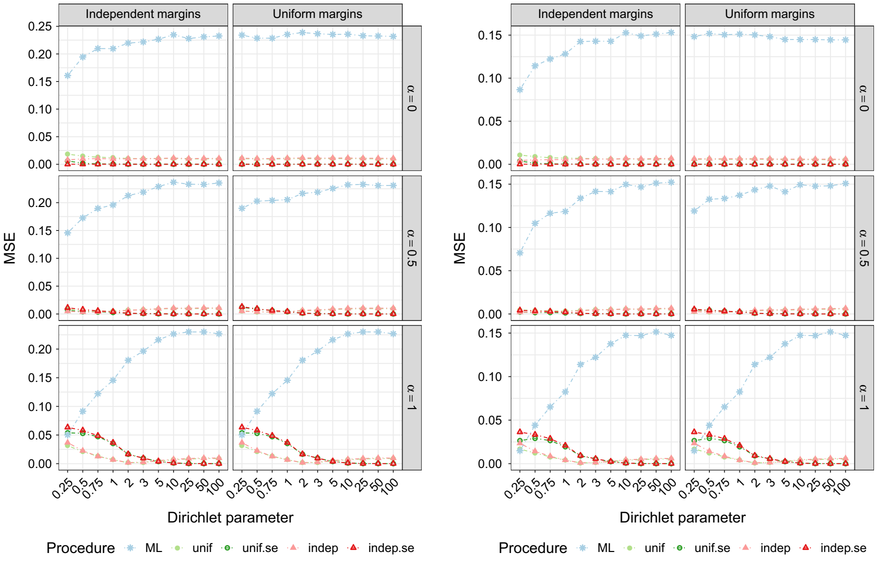
Fig. 3 The same scenario as in Fig. 2, but now MSE calculated for conditional mutual information $I(X, Y|Z)$



(a) X and Y have $k = 10$ levels, the number of observations is $n = 100$.

(b) X, Y and Z have $k = 5$ levels, the number of observations is $n = 125$.

Fig. 4 MSE calculated for estimated probabilities $p(x, y)$ (first panel) and $p(x, y, z)$ (second panel) for the model based on Dirichlet distribution



(a) X and Y have $k = 10$ levels, the number of observations is $n = 100$.

(b) X, Y and Z have $k = 5$ levels, the number of observations is $n = 125$.

Fig. 5 MSE calculated for mutual information $I(X, Y)$ (first panel) or conditional mutual information $I(X, Y|Z)$ (second panel) for the model based on Dirichlet distribution

equals Recall. Each experiment was repeated $N = 50$ times, the sample size was $n = 500$.

6.2.1 Results: variable selection for Bayesian Networks

Figure 6 shows that for many datasets the `indep.se` estimator works better than other procedures for all or almost all criteria. In the cases, where the differences between procedures are the most significant, the SE-based methods outperform `ML` and MSE-based estimators. It is also seen that one of the SE-regularized JMI methods (of the second or the third order) almost always wins. Surprisingly in some cases (`water`, `hailfinder`, `hepar2`) regularised JMI work much better than regularised JMI3. Thus application of higher order term approximations of CMI even when regularisation is applied, should be treated with caution, especially for small to moderate sample sizes. Overall, JMI with `indep.se` regularisation performs satisfactorily for all examples considered. We also note erratic behaviour of CIFE and CMIM which in some cases (see `pathfinder` and `win95pts`, respectively) perform poorly.

To gain deeper insight into the behaviour of the procedures in each step, we estimated TPR for $k \in \{1, 2, \ldots, 8\}$ most relevant variables with respect to the criterion. To obtain meaningful results, the target nodes with Markov Blankets greater or equal to 8 were chosen. In Fig. 7 we present results for four datasets. The datasets were chosen to show the behaviour of procedures in the cases where the differences showing superiority of SE-based methods are visible (`mildew` and `barley`) and where

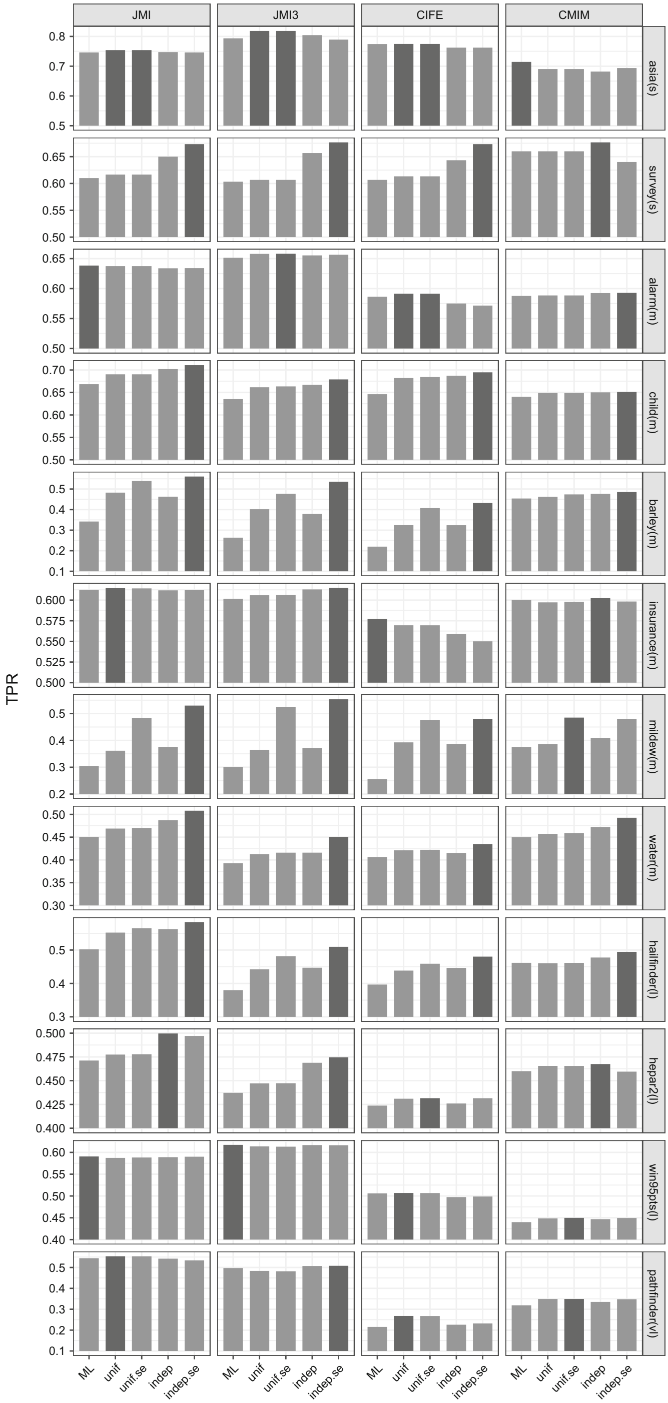


Fig. 6 Comparison of TPR across datasets averaged over target nodes. Letters in brackets stand for networks' sizes (*s* small, *m* medium, *l* large, *vl* very large). The winner's bar in each panel is marked with darker colour. Note that the vertical axis does not start at 0 and the starting point differs among rows

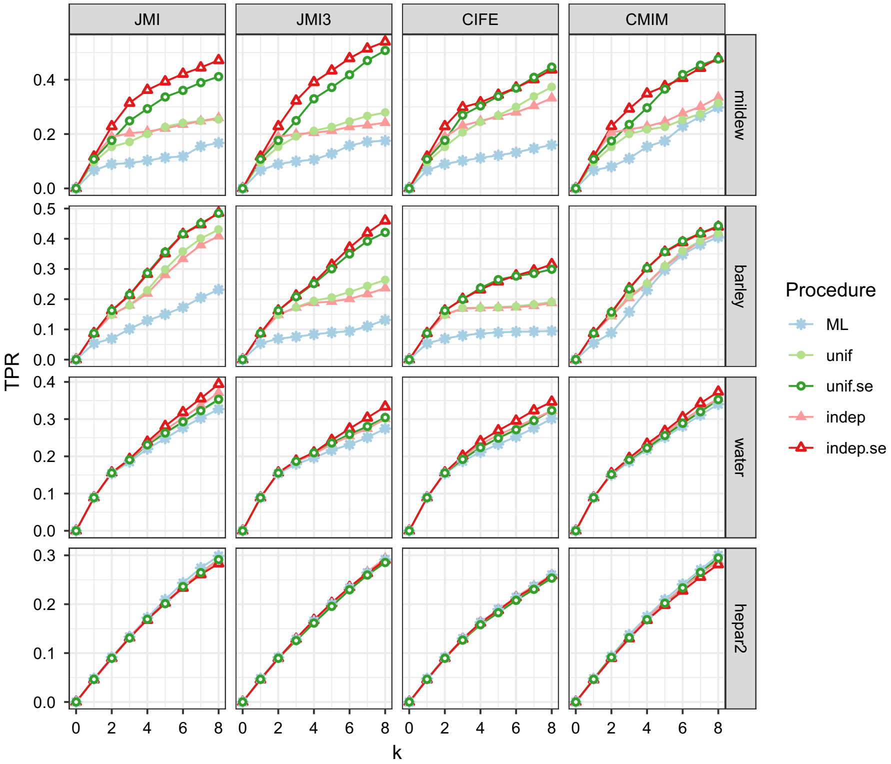


Fig. 7 True positive rate TPR for $k \in \{1, 2, \ldots, 8\}$ best variables with respect to the criterion and procedure for chosen datasets. The results were averaged over at least five target nodes for each dataset

there is only slight difference in performance with respect to procedures (`water`, `hepar2`). The results for other data sets investigated fall into one of these categories. The main advantage of shrinkage estimators shown in Fig. 7 is that they outperform the ML estimators even for small k. Also, for `mildew` and `barley` SE-based methods outperform MSE-based ones.

6.3 Classification example

The proposed methods of estimating probabilities yielding regularised versions of information criteria discussed above can be used for variable selection in classification problems. We present an example of such an use on a popular dataset `madelon` from the NIPS Feature Selection Challenge (cf. Guyon et al. 2003; training data consisting of 2000 observations was considered). Madelon contains 500 continuous variables and the label is binary. The dataset is split into three parts: a part for choosing variables (250 observations, the variables are discretized into 5 bins), a part for training (1225 observations; in our study the classifier is kNN with 5 neighbours) and a testing part (525 observations). The experiment is repeated 100 times, with a new random split

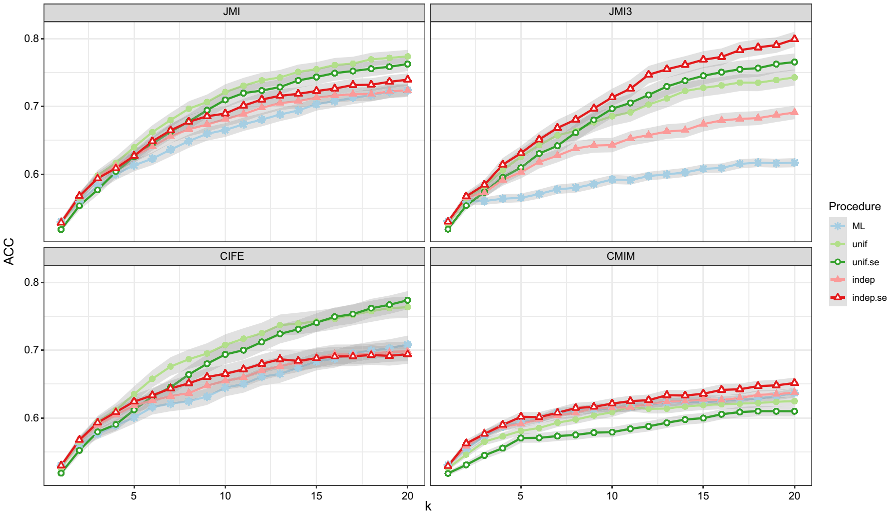


Fig. 8 Accuracy of nearest neighbour classifier with respect to the number of selected variables for `madelon`

for each repetition. The additional experiments for $n = 100$ and different number of bins equal to 3 and 7 and kNN parameter equal to 3, 7 are included in the supplement.

6.3.1 Results for **`madelon`** data set

Figure 8 shows the accuracy with respect to the number of selected variables. For JMI and JMI3 methods, the shrinkage procedures improve the performance of criteria, while for CIFE that effect is visible only for the procedures based on uniform distribution, whereas `indep` procedures behave similarly to `ML`. Among all methods considered JMI3 with `indep.se` regularisation works the best. Accuracy for CMIM, regardless of the regularisation applied, is low and that is the only criterion for which using one of the shrinkage estimators (`unif.se`) makes the result noticeably worse.

7 Conclusions

We have presented a new method of construction of the shrinkage estimator for the probability mass function based on SE minimisation which is reminiscent of bandwidth choice in density estimation based on minimisation of the Integrated SE. The method has an obvious theoretical motivation stemming from inequality (18) and is no more computationally demanding than its MSE-based counterpart. We have investigated properties of theoretically optimal regularisation parameters for the proposed and the previous methods and discussed crossvalidation

method of estimating the proposed parameter. In carefully designed numerical experiments we have shown that the relative performance of known regularisation methods depends on how much the marginal distributions deviate from the uniform distribution and that those methods are usually outperformed by SE-based ones with respect to the MSE. This empirically confirms inequality (18). Regularisation is also used in conjunction with information based selection criteria yielding promising results when compared to other methods for number of `bnlearn` data sets. In particular, JMI method with `indep.se` regularisation performed uniformly well for all considered data sets with respect to TPR.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

A Proof of Eq. (4)

Proof First recall the form of λ_{MSE} in (4):

$$\lambda_{MSE} = \frac{\sum_x \left(\mathrm{Var}(\hat{p}^{(2)}(x)) - \mathrm{Cov}(\hat{p}^{(1)}(x), \hat{p}^{(2)}(x))\right)}{\sum_x \mathbb{E}(\hat{p}^{(1)}(x) - \hat{p}^{(2)}(x))^2}.$$

Similarly to the proof of Lemma 1, we notice that $MSE(\lambda)$ is a quadratic function of λ:

$$\begin{aligned} MSE(\lambda) = \mathbb{E}\left(\sum_x (p(x) - \hat{p}_\lambda(x))^2\right) = \mathbb{E}\Bigg(\sum_x \Big[\lambda^2 \left(\hat{p}^{(1)}(x) - \hat{p}^{(2)}(x)\right)^2 \\ -2\lambda \left(p(x) - \hat{p}^{(2)}(x)\right)\left(\hat{p}^{(1)}(x) - \hat{p}^{(2)}(x)\right) + \left(p(x) - \hat{p}^{(2)}(x)\right)^2\Big]\Bigg). \end{aligned} \tag{37}$$

Therefore, we obtain

$$\lambda_{MSE} = \frac{\sum_x \mathbb{E}\left(p(x) - \hat{p}^{(2)}(x)\right)\left(\hat{p}^{(1)}(x) - \hat{p}^{(2)}(x)\right)}{\sum_x \mathbb{E}\left(\hat{p}^{(1)}(x) - \hat{p}^{(2)}(x)\right)^2} \tag{38}$$

and as $\mathbb{E}\,\hat{p}^{(2)}(x) = p(x)$, we have

$$\mathrm{Var}(\hat{p}^{(2)}(x)) = \mathbb{E}\left(\hat{p}^{(2)}(x) - p(x)\right)\hat{p}^{(2)}(x)$$

and

$$\mathrm{Cov}(\hat{p}^{(1)}(x), \hat{p}^{(2)}(x)) = \mathbb{E}\left(\hat{p}^{(2)}(x) - p(x)\right)\hat{p}^{(1)}(x).$$

□

B Proof of Theorem 1

Proof As $\hat{p}^{(2)}(x, y)$ is the maximum likelihood estimator of $p(x, y)$, its variance and second moment is well known and we omit the proof. We prove first corrected formula for $\mathbb{E}\left(p^{(1)}(x, y)\right)^2$. We have

$$\mathbb{E}\left(\hat{p}^{(1)}(x, y)\right)^2 = \frac{1}{n^4}\mathbb{E}\left[\sum_i \mathbb{I}(X_i = x)\sum_{i'} \mathbb{I}(X_{i'} = x)\sum_j \mathbb{I}(Y_j = y)\sum_{j'} \mathbb{I}(Y_{j'} = y)\right].$$

The contribution to the expected value of the summands on RHS depends on the number of elements of the set $A_1 = \{i, i', j, j'\}$.

- $|A_1| = 1$ $(i = j = i' = j')$
  The contribution is $np(x, y)/n^4$.
- $|A_1| = 2$
  We have four cases:

  1. $i = i', j = j', i \neq j$
     $n(n-1)p(x)p(y)/n^4$
  2. $i = j, i' = j', i \neq i'$ or $i = j', i = j', i \neq i'$. These two cases have the same contribution which equals:
     $n(n-1)p^2(x, y)/n^4$
  3. $i = i' = j, j \neq j'$ or $i = i' = j', j \neq j'$
     $n(n-1)p(x, y)p(y)/n^4$
  4. $i = j' = j, i \neq i'$ or $i' = j = j', i \neq i'$
     $n(n-1)p(x, y)p(x)/n^4$

- $|A| = 3$
  We have three cases:

  1. $i \neq i', j = j', j \neq i, j \neq i'$
     $n(n-1)(n-2)p^2(x)p(y)/n^4$
  2. $i = i', j \neq j', i \neq j, i \neq j'$
     $n(n-1)(n-2)p(x)p^2(y)/n^4$
  3. $i = j, i \neq i', j \neq j', i \neq j'$ (there are four cases like this—the remaining three have pairwise unequal indices except for one pair from $\{(i, j'), (i', j), (i', j')\}$)
     $n(n-1)(n-2)p(x, y)p(x)p(y)/n^4$

- $|A_1| = 4$
  $n(n-1)(n-2)(n-3)p^2(x)p^2(y)/n^4$

Summing up all of the terms we obtain the formula in Theorem 2. Similarly, we prove the formula for $\mathbb{E}(\hat{p}^{(1)}(x, y)\hat{p}^{(2)}(x, y))$:

$$\mathbb{E}(\hat{p}^{(1)}(x, y)\hat{p}^{(2)}(x, y)) = \frac{1}{n^3}\mathbb{E}\left[\sum_i \mathbb{I}(X_i = x)\sum_j \mathbb{I}(Y_j = y)\sum_k \mathbb{I}(X_k = x)\mathbb{I}(Y_k = y)\right].$$

We define $A_2 = \{i, j, k\}$, and we have three cases:

- $|A_2| = 1$
  The contribution is $np(x, y)/n^3$.
- $|A_2| = 2$
  We have three cases with corresponding contributions:

  1. $i = j, i \neq k$
     $n(n-1)p^2(x, y)/n^3$
  2. $i = k, i \neq j$
     $n(n-1)p(x)p(x, y)/n^3$
  3. $j = k, j \neq i$
     $n(n-1)p(y)p(x, y)/n^3$

- $|A_2| = 3$
  $n(n-1)(n-2)p(x)p(y)p(x, y)/n^3$

To obtain the formula for covariance of $\hat{p}^{(1)}(x, y)$ and $\hat{p}^{(2)}(x, y)$, we need to compute $\mathbb{E}\,\hat{p}^{(1)}(x, y)$ and then simply use

$$\mathrm{Cov}(\hat{p}^{(1)}(x, y), \hat{p}^{(2)}(x, y)) = \mathbb{E}\,\hat{p}^{(1)}(x, y)\hat{p}^{(2)}(x, y) - \mathbb{E}\,\hat{p}^{(1)}(x, y)\,\mathbb{E}\,\hat{p}^{(2)}(x, y).$$

To this end we note that

$$\begin{aligned}\mathbb{E}\,\hat{p}^{(1)}(x, y) &= \frac{1}{n^2}\,\mathbb{E}\left[\sum_i \mathbb{I}(X_i = x)\sum_j \mathbb{I}(Y_j = y)\right]\\ &= \left(np(x, y) + n(n-1)p(x)p(y)\right)/n^2.\end{aligned}$$

□

C Asymptotic behaviour of λ

We state the result analogous to Theorems 3 and 4 which study asymptotic behaviour of theoretical minimisers λ^U_{MSE} , λ^U_{SE}, λ^{Ind}_{MSE} and λ^{Ind}_{SE}. Note that in cases when the distribution is not uniform or is not a product of its marginals n times theoretical minimisers tend to the same limits as their empirical counterparts. The behaviour of λ^U_{MSE} and λ^U_{SE} in the uniform case is much simpler as they are equal to 1, whereas $\lambda^{Ind}_{MSE} \to 1$ in the independent case. The only unresolved case is the behaviour of λ^{Ind}_{SE} in the latter case. We include a partial result as the second part of (iv).

Theorem 5 *We have the following convergences provided $n \to \infty$:*

(i) *$n\lambda^U_{MSE} \to c$ when $p(x) \not\equiv 1/m$, c defined in (22) and $\lambda^U_{MSE} = 1$ otherwise;*

(ii) *$n\lambda^U_{SE} \to c$ a.e. when $p(x) \not\equiv 1/m$, c defined in (22), and $\lambda^U_{SE} = 1$ otherwise;*

(iii) *$n\lambda^{Ind}_{MSE} \to c$, when $p(x, y) \not\equiv p(x)p(y)$, c defined in (27) and $\lambda^{Ind}_{MSE} \to 1$ otherwise;*

(iv) *$n\,\lambda_{SE}^{Ind} \to c$ a.e. when $p(x,y) \not\equiv p(x)p(y)$, c defined in (27). Otherwise we have the following representation*

$$\hat\lambda_{SE}^{Ind} = \frac{\sum_{x,y}\begin{pmatrix}\left(\hat p^{(1)}(x,y) - \hat p^{(2)}(x,y)\right)^2 \\ +\left(p(x,y) - (\hat p^{(1)}(x,y)\right)\left(\hat p^{(1)}(x,y) - \hat p^{(2)}(x,y)\right)\end{pmatrix}}{\sum_{x,y}\left(\hat p^{(1)}(x,y) - \hat p^{(2)}(x,y)\right)^2}$$
$$=: \frac{M+R}{M}$$

and $\mathbb{E}\,R = 0$, $\mathbb{E}\,M = O(1/n)$

Proof We prove the second part of (iv) only as the proofs of the remaining parts are analogous but simpler than those of Theorems 3 and 4. The second part of (iii) follows from checking that the leading terms in the numerator and the denominator of λ_{MSE}^{Ind} (of order n^{-1}) are the same. We omit the details. In order to prove the second part of (iv) note that representation of λ_{SE}^{Ind} follows from a simple calculation. Moreover, using independence we have

$$\mathrm{Var}(\hat p^{(1)}(x,y)) = \left(\frac{n-1}{n}p^2(x) + \frac{1}{n}p(x)\right)\left(\frac{n-1}{n}p^2(y) + \frac{1}{n}p(y)\right) - p^2(x)p^2(y)$$

and it is seen that it coincides with $\mathrm{Cov}(p^{(1)}(x,y), \hat p^{(2)}(x,y))$ (see Theorem 1). Additionally, it is easily seen that

$$\mathrm{Var}(\hat p^{(2)}(x,y)) - \mathrm{Var}(\hat p^{(1)}(x,y)) \geq \frac{c_1}{n} \tag{39}$$

for some $c_1 > 0$. Thus we have

$$\mathbb{E}\,R = -\sum_{x,y}\mathrm{Var}(\hat p^{(1)}(x,y)) + \mathrm{Cov}(\hat p^{(1)}(x,y), \hat p^{(2)}(x,y)) = 0$$

whereas

$$\mathbb{E}\,M = \sum_{x,y}\mathbb{E}\left(\hat p^{(1)}(x,y) - \hat p^{(2)}(x,y)\right)^2 = \sum_{x,y}\Big(\mathrm{Var}(\hat p^{(1)}(x,y))$$
$$+\mathrm{Var}(\hat p^{(2)}(x,y)) - 2\mathrm{Cov}(\hat p^{(1)}(x,y), \hat p^{(2)}(x,y))\Big) \geq \frac{c_1}{n}$$

which ends the proof of the second part of (iv).

□

Remark 4 It follows from the proof of (iv) that in the case when X and Y are independent then

$$\mathrm{Var}(\hat p^{(2)}(x,y)) - \mathrm{Var}(\hat p_{\lambda_{SE}^{Ind}}(x,y)) \geq \frac{c_1}{n}$$

and thus shrinkage estimator has a smaller variance then ML estimator $\hat{p}^{(2)}$. This is intuitive as $\hat{p}^{(1)}$ is ML estimator in a smaller model assuming independence of X and Y and it has smaller variance then $\hat{p}^{(2)}$ [cf. (39)]. This property, which is likely to be preserved for approximate independence, is consistent with an aim of construction of James–Stein shrinkage estimators.

References

Agresti A (2013) Categorical data analysis, 3rd edn. Wiley, Hoboken

Bartoszyński R, Niewiadomska-Bugaj M (1996) Probability and statistical inference, 1st edn. Wiley, New York

Battiti R (1994) Using mutual information for selecting features in supervised neural-net learning. IEEE Trans Neural Netw 5(4):537–550

Borboudakis G, Tsamardinos I (2019) Forward–backward selection with early dropping. J Mach Learn Res 20:1–39

Brown G, Pocock A, Zhao MJ, Luján M (2012) Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. J Mach Learn Res 13(1):27–66

Cover TM, Thomas JA (2006) Elements of information theory. Wiley series in telecommunications and signal processing. Wiley-Interscience, New York

Fleuret F (2004) Fast binary feature selection with conditional mutual information. J Mach Learn Res 5:1531–1555

Guyon I (2003) Design of experiments for the NIPS 2003 variable selection benchmark. Presentation. www.nipsfsc.ecs.soton.ac.uk/papers/NIPS2003-Datasets.pdf

Hall P (1982) Limit theorems for stochastic measures of the accuracy of density estimators. Stoch Process Their Appl 13:11–25

Hall P (1983) Large sample optimality of least-squares crossvalidation in density estimation. Ann Stat 1:1156–1174

Hall P (1984) Central limit theorem for integrated square error of multivariate nonparametric density estimators. J Multivar Anal 14:1–16

Hall P, Marron J (1987) Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. Probab Theory Relat Fields 74:567–581. https://doi.org/10.1007/BF00363516

Hausser J, Strimmer K (2007) Entropy inference and the James–Stein estimator, with applications to non-linear gene association networks. J Mach Learn Res 10:1469–1484

Hausser J, Strimmer K (2014) Entropy: estimation of entropy, mutual information and related quantities. R package version 1.2.1. CRAN.R-project.org/package=entropy

James W, Stein C (1961) Estimation with quadratic loss. In: Proceedings of fourth Berkeley symposium on mathematical statistics and probability, pp 361–379

Kubkowski M, Mielniczuk J, Teisseyre P (2021) How to gain on power: novel conditional independence tests based on short expansion of conditional mutual information. J Mach Learn Res 22:1–57

Łazęcka M, Mielniczuk J (2020) Note on Machine Learning (2020) paper by Sechidis et al. Unpublished note

Ledoit O, Wolf M (2003) Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. J Empir Finance 10:603–621

Lewis D (1992) Feature selection and feature extraction for text categorisation. In: Proceedings of the workshop on speech and natural language

Lin D, Tang X (2006) Conditional infomax learning: an integrated framework for feature extraction and fusion. In: Proceedings of the 9th European conference on computer vision—Part I, ECCV'06, pp 68–82

Marron J, Härdle WK (1986) Random approximations to some measures of accuracy in nonparametric curve estimation. J Multivar Anal 20:91–113

Meyer P, Schretter C, Bontempi G (2008) Information-theoretic feature selection in microarray data using variable complementarity. IEEE Sel Top Signal Process 2(3):261–274

Mielniczuk J, Teisseyre P (2019) Stopping rules for mutual information-based feature selection. Neurocomputing 358:255–271

Nelsen R (2006) An introduction to copulas. Springer, New York

Pawluk M, Teisseyre P, Mielniczuk J (2019) Information-theoretic feature selection using high-order interactions. In: Machine learning, optimization, and data science. Springer, pp 51–63

Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(1):1226–1238

Rice J (1984) Bandwidth choice for regression estimation. Ann Stat 12(4):1215–1230

Schäffer I, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol. www.strimmerlab.org/publications/journals/shrinkcov2005.pdf

Scott D (2001) Parametric statistical modeling by minimum integrated square error. Technometrics 43:274–285

Scutari M (2010) Learning Bayesian networks with the bnlearn R package. J Stat Softw 35(3):1–22

Scutari M, Brogini A (2016) Bayesian structure learning with permutation tests. Commun Stat Theory Methods 41(16–17):3233–3243

Sechidis K, Azzimonti L, Pocock A, Corani G, Weatherall J, Brown G (2019) Efficient feature selection using shrinkage estimators. Mach Learn 108:1261–1286

Sechidis K, Azzimonti L, Pocock A, Corani G, Weatherall J, Brown G (2020) Corrigendum to: Efficient feature selection using shrinkage estimators. Mach Learn. https://doi.org/10.1007/s10994-020-05884-6

Stone C (1984) An asymptotically optimal window selection rule for kernel density estimates. Ann Stat 12(4):1285–1297

Sugiyama M, Kanamori T, Suzuki T, Plessis M, Liu S, Takeuchi I (2012) Density-difference estimation. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in neural information processing systems. Curran Associates, Inc

Vergara J, Estevez P (2014) A review of feature selection methods based on mutual information. Neural Comput Appl 24(1):175–186

Vinh N, Zhou S, Chan J, Bailey J (2016) Can high-order dependencies improve mutual information based feature selection? Pattern Recognit 53:45–58

Yang HH, Moody J (1999) Data visualization and feature selection: new algorithms for non-Gaussian data. Adv Neural Inf Process Syst 12:687–693

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.