Noname manuscript No.
(will be inserted by the editor)

Supplementary material: Efficient Feature Selection Using Shrinkage Estimators

Konstantinos Sechidis · Laura Azzimonti · Adam Pocock · Giorgio Corani · James Weatherall · Gavin Brown

A Proof of Theorems

A.1 Proof of Theorem 1

For this proof we will use Ledoit and Wolf theorem (Ledoit and Wolf, 2003), which derives an analytical expression for the optimal shrinkage intensity that guarantees minimal MSE. Using the fact that $\hat{p}^{\mathrm{ML}}(xy)$ is an unbiased estimator of $p(xy)$, the optimal shrinkage intensity takes the following form (Hausser and Strimmer, 2009, eq. (10)):

$$\lambda^* = \frac{\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\left(\mathrm{Var}\left[\hat{p}^{\mathrm{ML}}(xy)\right] - \mathrm{Cov}\left[\hat{p}^{\mathrm{ML}}(xy), \hat{p}^{\mathrm{Ind}}(xy)\right]\right)}{\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\left(\mathbb{E}\left[\left(\hat{p}^{\mathrm{ML}}(xy)\right)^2\right] + \mathbb{E}\left[\left(\hat{p}^{\mathrm{Ind}}(xy)\right)^2\right] - 2\mathbb{E}\left[\hat{p}^{\mathrm{ML}}(xy)\hat{p}^{\mathrm{Ind}}(xy)\right]\right)}$$

Following Hausser and Strimmer (2009) approach, we can derive a simple estimate of λ^* by replacing all variances, covariances and expectations with their empirical counterparts (Schäfer and Strimmer, 2005, eq. (8)):

$$\hat{\lambda}^* = \frac{\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\left(\overbrace{\widehat{\mathrm{Var}}\left[\hat{p}^{\mathrm{ML}}(xy)\right]}^{\texttt{First term}} - \overbrace{\widehat{\mathrm{Cov}}\left[\hat{p}^{\mathrm{ML}}(xy), \hat{p}^{\mathrm{Ind}}(xy)\right]}^{\texttt{Second term}}\right)}{\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\left(\underbrace{\widehat{\mathbb{E}}\left[\left(\hat{p}^{\mathrm{ML}}(xy)\right)^2\right]}_{\texttt{Third term}} + \underbrace{\widehat{\mathbb{E}}\left[\left(\hat{p}^{\mathrm{Ind}}(xy)\right)^2\right]}_{\texttt{Fourth term}} - 2\underbrace{\widehat{\mathbb{E}}\left[\hat{p}^{\mathrm{ML}}(xy)\hat{p}^{\mathrm{Ind}}(xy)\right]}_{\texttt{Fifth term}}\right)} \quad (1)$$

To derive the expressions of the five terms we will assume a random vector $\mathbf{N}$, whose elements are the counts N_{xy}, which is distributed as a $|\mathcal{X}||\mathcal{Y}|$ Multinomial random variable. The parameters are N, the total number of observations, and $\mathbf{p}$, the vector of the true underlying probabilities $p(xy)$. Under this assumption we will derive estimates for the five terms of eq. (1).

Konstantinos Sechidis · Gavin Brown
E-mail: [konstantinos.sechidis,gavin.brown]@manchester.ac.uk
School of Computer Science, University of Manchester, Manchester, UK

Laura Azzimonti · Giorgio Corani
E-mail: [laura,giorgio]@idsia.ch
Istituto Dalle Molle di studi sull' Intelligenza Artificiale (IDSIA), Manno, Switzerland

Adam Pocock
E-mail: adam.pocock@oracle.com
Oracle Labs, Burlington, MA, USA

James Weatherall
E-mail: James.Weatherall@astrazeneca.com
Advanced Analytics Centre, Global Medicines Development, AstraZeneca, Cambridge, UK

• `First term`: The first term can be written as follows:

$$\mathrm{Var}\left[\hat{p}^{\mathrm{ML}}(xy)\right] = \mathrm{Var}\left[\frac{N_{xy}}{N}\right] = \frac{1}{N^2}\left(\mathbb{E}\left[N_{xy}^2\right] - \mathbb{E}\left[N_{xy}\right]^2\right) \tag{2}$$

Under the Multinomial modelling, the first two moments are (Mosimann, 1962):

$$\mathbb{E}\left[N_{xy}\right] = Np(xy) \tag{3}$$
$$\mathbb{E}\left[N_{xy}^2\right] = N(N-1)p(xy)^2 + Np(xy) \tag{4}$$

By substituting eqs (3) and (4) into eq. (2) we get:

$$\mathrm{Var}\left[\hat{p}^{\mathrm{ML}}(xy)\right] = \frac{1}{N^2}\left(N(N-1)p(xy)^2 + Np(xy) - N^2p(xy)^2\right) = \frac{p(xy)}{N}\left(1-p(xy)\right)$$

The above term can be estimated as

$$\widehat{\mathrm{Var}}\left[\hat{p}^{\mathrm{ML}}(xy)\right] = \frac{\hat{p}^{\mathrm{ML}}(xy)}{N}\left(1-\hat{p}^{\mathrm{ML}}(xy)\right)$$

• `Second term`: The covariance term can be written as follows:

$$\begin{aligned}\mathrm{Cov}\left[\hat{p}^{\mathrm{ML}}(xy), \hat{p}^{\mathrm{Ind}}(xy)\right] &= \mathbb{E}\left[\hat{p}^{\mathrm{ML}}(xy)\hat{p}^{\mathrm{Ind}}(xy)\right] - \mathbb{E}\left[\hat{p}^{\mathrm{ML}}(xy)\right]\mathbb{E}\left[\hat{p}^{\mathrm{Ind}}(xy)\right] \\ &= \frac{1}{N^3}\left(\mathbb{E}\left[N_{xy}N_xN_y\right] - \mathbb{E}\left[N_{xy}\right]\mathbb{E}\left[N_xN_y\right]\right)\end{aligned} \tag{5}$$

The expected value $\mathbb{E}\left[N_xN_y\right]$ can be calculated as follows:

$$\mathbb{E}\left[N_xN_y\right] = \sum_{x'\in\mathcal{X}}\sum_{y'\in\mathcal{Y}}\mathbb{E}\left[N_{xy'}N_{x'y}\right] = \sum_{x'\in\mathcal{X}}\sum_{\substack{y'\in\mathcal{Y}\\ y'\neq y}}\mathbb{E}\left[N_{xy'}N_{x'y}\right] + \sum_{\substack{x'\in\mathcal{X}\\ x'\neq x}}\mathbb{E}\left[N_{xy}N_{x'y}\right] + \mathbb{E}\left[N_{xy}^2\right] \tag{6}$$

Under the Multinomial modelling, the second order moments can be written as follows (Mosimann, 1962):

$$\mathbb{E}\left[N_{xy'}N_{x'y}\right] = N(N-1)p(xy')p(x'y) \tag{7}$$
$$\mathbb{E}\left[N_{xy}N_{x'y}\right] = N(N-1)p(xy)p(x'y) \tag{8}$$

By substituting eqs (4),(7) and (8) into eq. (6) we get:

$$\begin{aligned}\mathbb{E}\left[N_xN_y\right] &= N(N-1)\left(\sum_{x'\in\mathcal{X}}\sum_{\substack{y'\in\mathcal{Y}\\ y'\neq y}}p(xy')p(x'y) + \sum_{\substack{x'\in\mathcal{X}\\ x'\neq x}}p(xy)p(x'y) + p(xy)^2\right) + Np(xy) \\ &= N(N-1)\left(\sum_{\substack{y'\in\mathcal{Y}\\ y'\neq y}}p(xy')p(y) + \sum_{\substack{x'\in\mathcal{X}\\ x'\neq x}}p(xy)p(x'y) + p(xy)^2\right) + Np(xy) \\ &= N(N-1)\left(\left(p(x)-p(xy)\right)p(y) + p(xy)\left(p(y)-p(xy)\right) + p(xy)^2\right) + Np(xy)\end{aligned}$$

By simple computation, we obtain

$$\mathbb{E}\left[N_xN_y\right] = N(N-1)p(x)p(y) + Np(xy) \tag{9}$$

Now we will calculate the expected value $\mathbb{E}\left[N_{xy}N_xN_y\right]$. This expectation can be written as follows:

$$\mathbb{E}\left[N_{xy}N_xN_y\right] = \sum_{x'\in\mathcal{X}}\sum_{y'\in\mathcal{Y}}\mathbb{E}\left[N_{xy}N_{xy'}N_{x'y}\right]$$

$$= \sum_{\substack{x' \in \mathcal{X} \\ x' \neq x}} \sum_{\substack{y' \in \mathcal{Y} \\ y' \neq y}} \mathbb{E}\left[N_{xy} N_{xy'} N_{x'y}\right] + \sum_{\substack{x' \in \mathcal{X} \\ x' \neq x}} \mathbb{E}\left[N_{xy}^2 N_{x'y}\right] + \sum_{\substack{y' \in \mathcal{Y} \\ y' \neq y}} \mathbb{E}\left[N_{xy}^2 N_{xy'}\right] + \mathbb{E}\left[N_{xy}^3\right] \tag{10}$$

Under the Multinomial modelling, the third order moments can be written as follows (Mosimann, 1962):

$$\mathbb{E}\left[N_{xy} N_{xy'} N_{x'y}\right] = N^{(3)} p(xy) p(xy') p(x'y) \tag{11}$$

$$\mathbb{E}\left[N_{xy}^2 N_{x'y}\right] = N^{(3)} p(xy)^2 p(x'y) + N^{(2)} p(xy) p(x'y) \tag{12}$$

$$\mathbb{E}\left[N_{xy}^2 N_{xy'}\right] = N^{(3)} p(xy)^2 p(xy') + N^{(2)} p(xy) p(xy') \tag{13}$$

$$\mathbb{E}\left[N_{xy}^3\right] = N^{(3)} p(xy)^3 + 3N^{(2)} p(xy)^2 + N p(xy) \tag{14}$$

where $N^{(2)} = N(N-1)$ and $N^{(a)} = N(N-1)\cdots(N-a+1)$, for $a > 2$.

By substituting eqs (11) - (14) into eq. (10) we get:

$$\begin{aligned}
\mathbb{E}\left[N_{xy} N_x N_y\right] = & \sum_{\substack{x' \in \mathcal{X} \\ x' \neq x}} \sum_{\substack{y' \in \mathcal{Y} \\ y' \neq y}} N^{(3)} p(xy) p(xy') p(x'y) \\
& + \sum_{\substack{x' \in \mathcal{X} \\ x' \neq x}} \left(N^{(3)} p(xy)^2 p(x'y) + N^{(2)} p(xy) p(x'y) \right) \\
& + \sum_{\substack{y' \in \mathcal{Y} \\ y' \neq y}} \left(N^{(3)} p(xy)^2 p(xy') + N^{(2)} p(xy) p(xy') \right) \\
& + N^{(3)} p(xy)^3 + 3N^{(2)} p(xy)^2 + N p(xy) \Leftrightarrow \\
\mathbb{E}\left[N_{xy} N_x N_y\right] = & N^{(3)} p(xy) \left(p(x) - p(xy)\right) \left(p(y) - p(xy)\right) \\
& + N^{(3)} p(xy)^2 \left(p(y) - p(xy)\right) + N^{(2)} p(xy) \left(p(y) - p(xy)\right) \\
& + N^{(3)} p(xy)^2 \left(p(x) - p(xy)\right) + N^{(2)} p(xy) \left(p(x) - p(xy)\right) \\
& + N^{(3)} p(xy)^3 + 3N^{(2)} p(xy)^2 + N p(xy)
\end{aligned}$$

By simple computation, we obtain

$$\mathbb{E}[N_{xy} N_x N_y] = N(N-1)(N-2) p(xy) p(x) p(y) + N(N-1) p(xy) \left(p(y) + p(x) + p(xy)\right) + N p(xy) \tag{15}$$

Thus, by substituting eqs (3), (9) and (15) into eq. (5) we get:

$$\begin{aligned}
& \mathrm{Cov}\left[\hat{p}^{\mathrm{ML}}(xy), \hat{p}^{\mathrm{Ind}}(xy)\right] \\
& = \frac{1}{N^3} \Big(N(N-1)(N-2) p(xy) p(x) p(y) + N(N-1) p(xy) \left(p(y) + p(x) + p(xy)\right) + N p(xy) \\
& \quad - \left(N(N-1) p(x) p(y) + N p(xy)\right) N p(xy) \Big)
\end{aligned}$$

Again, by simple computation we obtain:

$$\mathrm{Cov}\left[\hat{p}^{\mathrm{ML}}(xy), \hat{p}^{\mathrm{Ind}}(xy)\right] = \frac{p(xy)}{N^2} \Big((N-1) \left(p(x) + p(y) - 2p(x)p(y)\right) + 1 - p(xy) \Big)$$

The above term can be estimated as

$$\widehat{\mathrm{Cov}}\left[\hat{p}^{\mathrm{ML}}(xy), \hat{p}^{\mathrm{Ind}}(xy)\right] = \frac{\hat{p}^{\mathrm{ML}}(xy)}{N^2} \Big((N-1) \Big(\hat{p}^{\mathrm{ML}}(x) + \hat{p}^{\mathrm{ML}}(y) - 2\hat{p}^{\mathrm{ML}}(x) \hat{p}^{\mathrm{ML}}(y) \Big) + 1 - \hat{p}^{\mathrm{ML}}(xy) \Big)$$

• `Third term`: This expectation term can be calculated as follows:

$$\mathbb{E}\left[\left(\hat{p}^{\mathrm{ML}}(xy)\right)^2\right] = \frac{1}{N^2}\mathbb{E}\left[N_{xy}^2\right]$$

Using eq. (4), the last expression can be written as

$$\mathbb{E}\left[\left(\hat{p}^{\mathrm{ML}}(xy)\right)^2\right] = \frac{p(xy)}{N}\left((N-1)p(xy)+1\right)$$

Finally, the above term can be estimated as

$$\widehat{\mathbb{E}}\left[(\hat{p}^{\mathrm{ML}}(xy))^2\right] = \frac{\hat{p}^{\mathrm{ML}}(xy)}{N}\Big(\,(N-1)\,\hat{p}^{\mathrm{ML}}(xy)+1\Big)$$

• `Fourth term`: This term can be written as follows:

$$\begin{aligned}\mathbb{E}\left[(\hat{p}^{\mathrm{Ind}}(xy))^2\right] &= \frac{\mathbb{E}\left[N_x^2 N_y^2\right]}{N^4} = \frac{\mathbb{E}\left[\left(\sum_{y\in\mathcal{Y}} N_{xy}\right)^2\left(\sum_{x\in\mathcal{X}} N_{xy}\right)^2\right]}{N^4} \\ &= \frac{\sum_{x',x''\in\mathcal{X}}\sum_{y',y''\in\mathcal{Y}}\mathbb{E}\left[N_{xy'}N_{xy''}N_{x'y}N_{x''y}\right]}{N^4} \end{aligned} \tag{16}$$

According the known fourth order moments formulas (Mosimann, 1962), we need to treat the terms in eq. (16) by splitting them in five categories:

A. four different terms;
B. three different terms over four;
C. two different terms over four (two couples equal);
D. two different terms over four (a triplet and a single element);
E. a single element to the power four.

The sum of forth order moments can thus be written as the sum of five different terms, i.e.,

$$\sum_{x',x''\in\mathcal{X}}\sum_{y',y''\in\mathcal{Y}}\mathbb{E}\left[N_{xy'}N_{xy''}N_{x'y}N_{x''y}\right] = A(xy)+B(xy)+C(xy)+D(xy)+E(xy),$$

where

$$\begin{aligned}
A(xy) =& \sum_{\substack{x',x''\in\mathcal{X}\\ x'\neq x''\neq x}}\sum_{\substack{y',y''\in\mathcal{Y}\\ y'\neq y''}}\mathbb{E}\left[N_{xy'}N_{xy''}N_{x'y}N_{x''y}\right] + 2\sum_{\substack{x'\in\mathcal{X}\\ x'\neq x}}\sum_{\substack{y',y''\in\mathcal{Y}\\ y'\neq y''\neq y}}\mathbb{E}\left[N_{xy'}N_{xy''}N_{x'y}N_{xy}\right], \\
B(xy) =& \sum_{\substack{x',x''\in\mathcal{X}\\ x'\neq x''\neq x}}\sum_{y'\in\mathcal{Y}}\mathbb{E}\left[N_{xy'}^2N_{x'y}N_{x''y}\right] + 2\sum_{\substack{x'\in\mathcal{X}\\ x'\neq x}}\sum_{\substack{y'\in\mathcal{Y}\\ y'\neq y}}\mathbb{E}\left[N_{xy'}^2N_{x'y}N_{xy}\right] \\
&+ \sum_{x'\in\mathcal{X}}\sum_{\substack{y',y''\in\mathcal{Y}\\ y'\neq y''}}\mathbb{E}\left[N_{xy'}N_{xy''}N_{x'y}^2\right] + 2\sum_{\substack{x'\in\mathcal{X}\\ x'\neq x}}\sum_{\substack{y'\in\mathcal{Y}\\ y'\neq y}}\mathbb{E}\left[N_{xy'}N_{xy}N_{x'y}^2\right] \\
&+ 4\sum_{\substack{x'\in\mathcal{X}\\ x'\neq x}}\sum_{\substack{y'\in\mathcal{Y}\\ y'\neq y}}\mathbb{E}\left[N_{xy'}N_{xy}^2N_{x'y}\right], \\
C(xy) =& \sum_{x'\in\mathcal{X}}\sum_{\substack{y'\in\mathcal{Y}\\ y'\neq y}}\mathbb{E}\left[N_{xy'}^2N_{x'y}^2\right] + \sum_{\substack{x'\in\mathcal{X}\\ x'\neq x}}\mathbb{E}\left[N_{xy}^2N_{x'y}^2\right], \\
D(xy) =& 2\sum_{\substack{x'\in\mathcal{X}\\ x'\neq x}}\mathbb{E}\left[N_{xy}^3N_{x'y}\right] + 2\sum_{\substack{y'\in\mathcal{Y}\\ y'\neq y}}\mathbb{E}\left[N_{xy}^3N_{xy'}\right],
\end{aligned}$$

$$E(xy) = \mathbb{E}\left[N_{xy}^4\right].$$

Under the Multinomial modelling, the fourth order moments that we need to calculate these five terms can be written as follows (Mosimann, 1962):

$$\mathbb{E}\left[N_{xy'}N_{xy''}N_{x'y}N_{x''y}\right] = N^{(4)}p(xy')p(xy'')p(x'y)p(x''y) \tag{17}$$

$$\mathbb{E}\left[N_{xy'}N_{xy''}N_{x'y}N_{xy}\right] = N^{(4)}p(xy')p(xy'')p(x'y)p(xy) \tag{18}$$

$$\mathbb{E}\left[N_{xy'}^2N_{x'y}N_{x''y}\right] = N^{(4)}p(xy')^2p(x'y)p(x''y) + N^{(3)}p(xy')p(x'y)p(x''y) \tag{19}$$

$$\mathbb{E}\left[N_{xy'}^2N_{x'y}N_{xy}\right] = N^{(4)}p(xy')^2p(x'y)p(xy) + N^{(3)}p(xy')p(x'y)p(xy) \tag{20}$$

$$\mathbb{E}\left[N_{xy'}N_{xy''}N_{x'y}^2\right] = N^{(4)}p(xy')p(xy'')p(x'y)^2 + N^{(3)}p(xy')p(xy'')p(x'y) \tag{21}$$

$$\mathbb{E}\left[N_{xy'}N_{xy}N_{x'y}^2\right] = N^{(4)}p(xy')p(xy)p(x'y)^2 + N^{(3)}p(xy')p(xy)p(x'y) \tag{22}$$

$$\mathbb{E}\left[N_{xy'}N_{xy}^2N_{x'y}\right] = N^{(4)}p(xy')p(xy)^2p(x'y) + N^{(3)}p(xy')p(xy)p(x'y) \tag{23}$$

$$\mathbb{E}\left[N_{xy'}^2N_{x'y}^2\right] = N^{(4)}p(xy')^2p(x'y)^2 + N^{(3)}p(xy')p(x'y)\left(p(xy') + p(x'y)\right) + N^{(2)}p(xy')p(x'y) \tag{24}$$

$$\mathbb{E}\left[N_{xy}^2N_{x'y}^2\right] = N^{(4)}p(xy)^2p(x'y)^2 + N^{(3)}p(xy)p(x'y)\left(p(xy) + p(x'y)\right) + N^{(2)}p(xy)p(x'y) \tag{25}$$

$$\mathbb{E}\left[N_{xy}^3N_{x'y}\right] = N^{(4)}p(xy)^3p(x'y) + 3N^{(3)}p(xy)^2p(x'y) + N^{(2)}p(xy)p(x'y) \tag{26}$$

$$\mathbb{E}\left[N_{xy}^3N_{xy'}\right] = N^{(4)}p(xy)^3p(xy') + 3N^{(3)}p(xy)^2p(xy') + N^{(2)}p(xy)p(xy') \tag{27}$$

$$\mathbb{E}\left[N_{xy}^4\right] = N^{(4)}p(xy)^4 + 6N^{(3)}p(xy)^3 + 7N^{(2)}p(xy)^2 + Np(xy). \tag{28}$$

Using eqs (17) - (28) and some simple computation, we obtain the following expressions for the five terms

$$\begin{aligned}
A(xy) =& N^{(4)}\Bigg[\Bigg(p(x)^2 - \sum_{y'\in\mathcal{Y}} p(xy')^2\Bigg)\Bigg(p(y)^2 - \sum_{x'\in\mathcal{X}} p(x'y)^2\Bigg) \\
&-4\big(p(x) - p(xy)\big)p(xy)^2\big(p(y) - p(xy)\big)\Bigg], \\
B(xy) =& N^{(4)}\Bigg[\sum_{y'\in\mathcal{Y}} p(xy')^2\Bigg(p(y)^2 - \sum_{x'\in\mathcal{X}} p(x'y)^2\Bigg) - 2p(xy)^3(p(y) - p(xy)) \\
&+ \sum_{x'\in\mathcal{X}} p(x'y)^2\Bigg(p(x)^2 - \sum_{y'\in\mathcal{Y}} p(xy')^2\Bigg) - 2p(xy)^3(p(x) - p(xy)) \\
&+ 4(p(x) - p(xy))p(xy)^2(p(y) - p(xy))\Bigg] \\
&+ N^{(3)}\Bigg[p(x)\Bigg(p(y)^2 - \sum_{x'\in\mathcal{X}} p(x'y)^2\Bigg) - 2p(xy)^2(p(y) - p(xy)) \\
&+ p(y)\Bigg(p(x)^2 - \sum_{y'\in\mathcal{Y}} p(xy')^2\Bigg) - 2p(xy)^2(p(x) - p(xy)) \\
&+ 4(p(x) - p(xy))p(xy)(p(y) - p(xy))\Bigg] \\
C(xy) + E(xy) =& N^{(4)} \sum_{x'\in\mathcal{X}} p(x'y)^2 \sum_{y'\in\mathcal{Y}} p(xy')^2 \\
&+ N^{(3)}\Bigg[4p(xy)^3 + p(x)\sum_{x'\in\mathcal{X}} p(x'y)^2 + p(y)\sum_{y'\in\mathcal{X}} p(xy')^2\Bigg]
\end{aligned}$$

$$
\begin{aligned}
&+ N^{(2)}\left[6p(xy)^2 + p(x)p(y)\right], \\
D(xy) =& 2N^{(4)}\left[p(xy)^3(p(x)+p(y)-2p(xy)\right] \\
&+ 6N^{(3)}p(xy)^2(p(x)+p(y)-2p(xy)) \\
&+ 2N^{(2)}p(xy)(p(x)+p(y)-2p(xy)).
\end{aligned}
$$

By simple computation, we obtain

$$
\begin{aligned}
\sum_{x',x''\in\mathcal{X}}\sum_{y',y''\in\mathcal{Y}} \mathbb{E}\left[N_{xy'}N_{xy''}N_{x'y}N_{x''y}\right] =& N^{(4)}p(x)^2p(y)^2 \\
&+N^{(3)}p(x)p(y)(p(x)+p(y)+4p(xy)) \\
&+N^{(2)}\left[2p(xy)(p(x)+p(y))+2p(xy)^2+p(x)p(y)\right] \\
&+Np(xy).
\end{aligned}
$$

Since the parameters $p(xy)$ are unknown, we substitute the ML estimates in order to estimate the second moment of $\hat{p}^{\text{Ind}}(xy)$, i.e.,

$$
\begin{aligned}
\widehat{\mathbb{E}}\left[(\hat{p}^{\text{Ind}}(xy))^2\right] = \frac{1}{N^3}\Big(& (N-1)(N-2)(N-3)\big((\hat{p}^{\text{ML}}(x)\hat{p}^{\text{ML}}(y))^2 \\
&+(N-1)(N-2)\hat{p}^{\text{ML}}(x)\hat{p}^{\text{ML}}(y)\big((\hat{p}^{\text{ML}}(x)+\hat{p}^{\text{ML}}(y)+4\hat{p}^{\text{ML}}(xy))\big) \\
&+(N-1)\big(2\hat{p}^{\text{ML}}(xy)(\hat{p}^{\text{ML}}(x)+\hat{p}^{\text{ML}}(y))+2(\hat{p}^{\text{ML}}(xy))^2+\hat{p}^{\text{ML}}(x)\hat{p}^{\text{ML}}(y)\big)+\hat{p}^{\text{ML}}(xy)\Big).
\end{aligned}
$$

• `Fifth term`: This expectation term can be calculated as follows:

$$
\mathbb{E}\left[\hat{p}^{\text{ML}}(xy)\hat{p}^{\text{Ind}}(xy)\right] = \frac{1}{N^3}\mathbb{E}\left[N_{xy}N_xN_y\right]
$$

Using eq. (15), the last expression can be written as

$$
\mathbb{E}\left[\hat{p}^{\text{ML}}(xy)\hat{p}^{\text{Ind}}(xy)\right] = \frac{p(xy)}{N^2}\left((N-1)(N-2)p(x)p(y)+(N-1)\left(p(y)+p(x)+p(xy)\right)+1\right)
$$

Finally, the above term can be estimated as

$$
\widehat{\mathbb{E}}\left[\hat{p}^{\text{ML}}(xy)\hat{p}^{\text{Ind}}(xy)\right]=\frac{\hat{p}^{\text{ML}}(xy)}{N^2}\left((N-1)(N-2)\hat{p}^{\text{ML}}(x)\hat{p}^{\text{ML}}(y)+(N-1)\left(\hat{p}^{\text{ML}}(x)+\hat{p}^{\text{ML}}(y)+\hat{p}^{\text{ML}}(xy)\right)+1\right)
$$

A.2 Proof of Theorem 2

Let us re-express CMI criterion, presented in main text's eq. (10), using the identity $I(A;B|C)-I(A;B)=I(A;C|B)-I(A;C)$:

$$
J_{\text{CMI}}(X_k) = I(X_k;Y) - I(X_k;\mathbf{X}_\theta) + I(X_k;\mathbf{X}_\theta|Y). \tag{29}
$$

Now we will decompose the second and third term, which capture redundancy and complementarity, using Assumptions 1 and 2 respectively.

Redundancy term - This term can be written as:

$$
I(X_k;\mathbf{X}_\theta) = H(\mathbf{X}_\theta) - H(\mathbf{X}_\theta|X_k)
$$

Using Assumption 1 we can re-write the rhs as:

$$I(X_k;\mathbf{X}_\theta) = H(\mathbf{X}_\theta) - \sum_{X_j\in\mathbf{X}_\theta} H(X_j|X_k) - \sum_{X_j\in\mathbf{X}_\theta}\sum_{\substack{X_i\in\mathbf{X}_\theta\\ i\neq j}} H(X_i|X_kX_j)$$

Using the identity $H(A|B) = H(A) - I(B;A)$ for the second term, and $H(A|BC) = H(A|C) - I(B;A|C)$ for the third term, the above equation can be written as:

$$\begin{aligned} I(X_k;\mathbf{X}_\theta) = H(\mathbf{X}_\theta) &- \sum_{X_j\in\mathbf{X}_\theta} H(X_j) + \sum_{X_j\in\mathbf{X}_\theta} I(X_k;X_j) \\ &- \sum_{X_j\in\mathbf{X}_\theta}\sum_{\substack{X_i\in\mathbf{X}_\theta\\ i\neq j}} H(X_i|X_j) + \sum_{X_j\in\mathbf{X}_\theta}\sum_{\substack{X_i\in\mathbf{X}_\theta\\ i\neq j}} I(X_k;X_i|X_j) \end{aligned} \tag{30}$$

Complementarity term - This term can be written as:

$$I(X_k;\mathbf{X}_\theta|Y) = H(\mathbf{X}_\theta|Y) - H(\mathbf{X}_\theta|X_kY)$$

Using Assumption 2 we can re-write the rhs as:

$$I(X_k;\mathbf{X}_\theta|Y) = H(\mathbf{X}_\theta|Y) - \sum_{X_j\in\mathbf{X}_\theta} H(X_j|X_kY) - \sum_{X_j\in\mathbf{X}_\theta}\sum_{\substack{X_i\in\mathbf{X}_\theta\\ i\neq j}} H(X_i|X_kX_jY)$$

Using the identity $H(A|BC) = H(A|C) - I(B;A|C)$ for the second term, and $H(A|BCD) = H(A|CD) - I(B;A|CD)$ for the third term, the above equation can be written as:

$$\begin{aligned} I(X_k;\mathbf{X}_\theta|Y) = H(\mathbf{X}_\theta|Y) &- \sum_{X_j\in\mathbf{X}_\theta} H(X_j|Y) + \sum_{X_j\in\mathbf{X}_\theta} I(X_k;X_j|Y) \\ &\sum_{X_j\in\mathbf{X}_\theta}\sum_{\substack{X_i\in\mathbf{X}_\theta\\ i\neq j}} H(X_i|X_jY) + \sum_{X_j\in\mathbf{X}_\theta}\sum_{\substack{X_i\in\mathbf{X}_\theta\\ i\neq j}} I(X_k;X_i|X_jY) \end{aligned} \tag{31}$$

The derived redundancy and complementarity terms, eqs. (30) and (31), can be substituted to eq. (29) so the CMI criterion can be written as follows:

$$\begin{aligned} J''_{\text{CMI}}(X_k) = I(X_k;Y) &- \sum_{X_j\in\mathbf{X}_\theta} I(X_k;X_j) + \sum_{X_j\in\mathbf{X}_\theta} I(X_k;X_j|Y) \\ &- \sum_{X_j\in\mathbf{X}_\theta}\sum_{\substack{X_i\in\mathbf{X}_\theta\\ i\neq j}} I(X_i;X_k|X_j) + \sum_{X_j\in\mathbf{X}_\theta}\sum_{\substack{X_i\in\mathbf{X}_\theta\\ i\neq j}} I(X_i;X_k|X_jY) + \text{Const}, \end{aligned}$$

where Const are constant terms with respect to X_k, as such removing them will have no effect on the choice of feature. Removing these terms we have an equivalent criterion.

A.3 Proof of Theorem 3

The JMI-3 criterion can be written in the following way:

$$J_{\text{JMI-3}}(X_k) = \sum_{X_j\in\mathbf{X}_\theta}\sum_{\substack{X_i\in\mathbf{X}_\theta\\ i\neq j}} I(X_kX_iX_j;Y) = \sum_{X_j\in\mathbf{X}_\theta}\sum_{\substack{X_i\in\mathbf{X}_\theta\\ i\neq j}} (I(X_iX_j;Y) + I(X_k;Y|X_iX_j))$$

The first term is constant with respect to X_k, as such removing it will have no effect on the choice of feature. We will use the identity $I(A;B|C) = I(A;B) - I(A;C) + I(A;C|B)$ to re-express the conditional mutual information term

$$J_{\text{JMI-3}}(X_k) \propto \sum_{X_j \in \mathbf{X}_\theta} \sum_{\substack{X_i \in \mathbf{X}_\theta \\ i \neq j}} (I(X_k;Y) - I(X_k;X_iX_j) + I(X_k;X_iX_j|Y))$$

The last two terms of the rhs can written as follows: $I(X_k;X_iX_j) = I(X_k;X_j) + I(X_k;X_i|X_j)$ and $I(X_k;X_iX_j|Y) = I(X_k;X_j|Y) + I(X_k;X_i|X_jY)$. Thus the JMI-3 criterion is

$$\begin{aligned}
J_{\text{JMI-3}}(X_k) \propto & \sum_{X_j \in \mathbf{X}_\theta} \sum_{\substack{X_i \in \mathbf{X}_\theta \\ i \neq j}} \Big(I(X_k;Y) - I(X_k;X_j) - I(X_k;X_i|X_j) + I(X_k;X_j|Y) + I(X_k;X_i|X_jY) \Big) = \\
= & |\mathbf{X}_\theta|(|\mathbf{X}_\theta|-1) I(X_k;Y) - (|\mathbf{X}_\theta|-1) \sum_{X_j \in \mathbf{X}_\theta} I(X_k;X_j) - \sum_{X_j \in \mathbf{X}_\theta} \sum_{\substack{X_i \in \mathbf{X}_\theta \\ i \neq j}} I(X_k;X_i|X_j) \\
& + (|\mathbf{X}_\theta|-1) \sum_{X_j \in \mathbf{X}_\theta} I(X_k;X_j|Y) + \sum_{X_j \in \mathbf{X}_\theta} \sum_{\substack{X_i \in \mathbf{X}_\theta \\ i \neq j}} I(X_k;X_i|X_jY) \\
\propto & I(X_k;Y) - \frac{1}{|\mathbf{X}_\theta|} \sum_{X_j \in \mathbf{X}_\theta} I(X_k;X_j) + \frac{1}{|\mathbf{X}_\theta|} \sum_{X_j \in \mathbf{X}_\theta} I(X_k;X_j|Y) \\
& - \frac{1}{|\mathbf{X}_\theta|(|\mathbf{X}_\theta|-1)} \sum_{X_j \in \mathbf{X}_\theta} \sum_{\substack{X_i \in \mathbf{X}_\theta \\ i \neq j}} I(X_k;X_i|X_j) + \frac{1}{|\mathbf{X}_\theta|(|\mathbf{X}_\theta|-1)} \sum_{X_j \in \mathbf{X}_\theta} \sum_{\substack{X_i \in \mathbf{X}_\theta \\ i \neq j}} I(X_k;X_i|X_jY)
\end{aligned}$$

The last expression shows that the JMI-3 criterion can be decomposed in the five terms of main text's eq. (17) with coefficients: $\beta = \gamma = 1/|\mathbf{X}_\theta|$, and $\beta' = \gamma' = 1/|\mathbf{X}_\theta|(|\mathbf{X}_\theta| - 1)$.

A.4 Proof of Theorem 4

Using identity $I(A;B|C) = I(A;B) - I(A;C) + I(A;C|B)$, we can re-express the CMIM-3 criterion in the following way:

$$\begin{aligned}
J_{\text{CMIM-3}}(X_k) &= \min_{\substack{X_j \in \mathbf{X}_\theta \\ X_i \in \mathbf{X}_\theta \\ i \neq j}} \Big[I(X_k;Y) - I(X_k;X_iX_j) + I(X_k;X_iX_j|Y) \Big] \\
&= I(X_k;Y) - \max_{\substack{X_j \in \mathbf{X}_\theta \\ X_i \in \mathbf{X}_\theta \\ i \neq j}} \Big[I(X_k;X_iX_j) - I(X_k;X_iX_j|Y) \Big]
\end{aligned}$$

The last two terms of the rhs can written as follows: $I(X_k;X_iX_j) = I(X_k;X_j) + I(X_k;X_i|X_j)$ and $I(X_k;X_iX_j|Y) = I(X_k;X_j|Y) + I(X_k;X_i|X_jY)$. Thus the CMIM-3 criterion can be decomposed as follows:

$$J_{\text{CMIM-3}}(X_k) = I(X_k;Y) - \max_{\substack{X_j \in \mathbf{X}_\theta \\ X_i \in \mathbf{X}_\theta \\ i \neq j}} \Big[I(X_k;X_j) - I(X_k;X_j|Y) + I(X_k;X_i|X_j) - I(X_k;X_i|X_jY) \Big]$$

B Protocol for generating synthetic data

B.1 Generating data for Section 3.2

The data are generated as follows:

1. We specify the desired values for the population MI i.e. $I(X;Y) \in (0, 0.05]$
2. We generate the population values for the probabilities $p(xy)$ by sampling from a Dirichlet distribution. Using these probabilities we can calculate the population value of the mutual information $I(X;Y)$, and if it is inside the desirable values (specified in Step 1) we keep the probabilities, else we sample again. Every time we sample from a Dirichlet, Dir(α), the concentration parameter α is a number chosen randomly from: $[0.3 - 3]$. It is interesting to mention here that the concentration parameter α controls the concentration of the prior: large α provides uniform distributions, and as a result small MI values, while small values lead to more concentrated distributions, which generate higher mutual information values. Using the probabilities $p(xy)$, we calculate $p(x) = \sum_{y \in \mathcal{Y}} p(xy)$, and $p(y|x) = \frac{p(xy)}{p(x)}$.
3. We generate the values of X,$\{x^n\}_{n=1}^N$, by sampling N data from the marginal distribution $p(x)$. Then we generate the values of Y by sampling N data from the conditional distributions $\{p(y|X = x^n)\}_{n=1}^N$.

B.2 Generating data for Section 3.3

The data are generated as follows:

1. We specify the desired values for the population CMI i.e. $I(X;Y|Z) \in (0, 0.05]$
2. We generate the population values for the probabilities $p(x)$ and $p(y)$ by sampling from a Dirichlet distribution, and for each value of X and Y we generate the values of $p(z|xy)$ by sampling again from a Dirichlet. Using these probabilities we can calculate the population value of the conditional mutual information $I(X;Y|Z)$, and if it is inside the desirable values (specified in Step 1) we keep the probabilities, else we sample again. Every time we sample from a Dirichlet, Dir(α) the concentration parameter α is a number chosen randomly from: $[0.3 - 3]$.
3. We generate the values of X and Y, $\{x^n y^n\}_{i=n}^N$, by sampling N data from the marginal distribution $p(x)$ and $p(y)$ respectively. Then we generate the values of Z by sampling n data from the conditional distributions $\{p(z|X = x^n, Y = y^n)\}_{n=1}^N$.

C Datasets Used

Table 1 shows the characteristics of the 11 benchmark BN[1] that we used to simulate the data in Section 5, while Table 2 shows the details of the 20 UCI datasets[2] used in Section 6.

Table 1 Summary of benchmark Bayesian networks

#	Network	Total number of nodes	Number of target nodes	Average MB size of target nodes
1.	asia	8	4	3.50
2.	child	20	8	5
3.	insurance	27	19	6.05
4.	water	32	16	10.25
5.	alarm	37	12	5.42
6.	barley2	48	29	6.48
7.	hailfinder	56	24	5.04
8.	hepar2	70	16	11
9.	win95pts	76	25	7.76
10.	pathfinder	109	30	5.87
11.	andes	223	112	7.32

[1] Downloaded from http://www.bnlearn.com/bnrepository/

[2] downloaded from https://archive.ics.uci.edu/ml/datasets.html

Table 2 Summary of UCI datasets.

#	Name	Examples	Features	Classes	#	Name	Examples	Features	Classes
1.	lungcancer	32	56	3	11.	breast	569	30	2
2.	soybean	47	35	4	12.	pima	768	8	2
3.	wine	178	13	3	13.	semeion	1593	256	10
4.	parkinsons	195	22	2	14.	splice	3175	60	3
5.	sonar	208	60	2	15.	krvskp	3196	36	2
6.	spect	267	22	2	16.	spambase	4601	57	2
7.	heart	270	13	2	17.	waveform	5000	40	3
8.	liver	345	6	2	18.	landsat	6435	36	6
9.	ionosphere	351	33	2	19.	musk2	6598	166	2
10.	congress	435	16	2	20.	mushroom	8124	21	2

D Experimental Results

Tables 3 and 4 contain the complete results of Section 5 for each dataset. Table 5 contain the results presented in Section 6.

Table 3 Comparison between our suggested high-order FS criteria in terms of their ability to identify the correct features (TPR) for BN with: (a) Sample size = 500 , and (b) Sample size = 2500. The best method (i.e. highest TPR) is highlighted with bold font and at the bottom of the table we present the average ranking score of each method across all datasets.

(a) TPR with 500 sample size

	$J^{\text{Ind-JS}}_{\text{JMI-3}}$	$J^{\text{Ind-JS}}_{\text{JMI-4}}$	$J^{\text{Ind-JS}}_{\text{CMIM-3}}$	$J^{\text{Ind-JS}}_{\text{CMIM-4}}$
asia	0.798	0.798	**0.808**	0.758
child	**0.773**	0.741	0.661	0.647
hailfinder	**0.497**	0.486	0.450	0.446
alarm	0.709	**0.733**	0.618	0.671
pathfinder	**0.450**	0.428	0.412	0.373
insurance	**0.634**	0.611	0.618	0.633
barley2	**0.479**	0.444	0.460	0.408
andes	**0.591**	0.576	0.527	0.527
win95pts	**0.600**	0.578	0.477	0.445
water	**0.507**	0.492	0.433	0.388
hepar2	**0.501**	0.489	0.494	0.463
Avg ranking	**1.227**	2.318	2.727	3.727
CD diagram		Figure 5(a)		

(b) TPR with 2500 sample size

	$J^{\text{Ind-JS}}_{\text{JMI-3}}$	$J^{\text{Ind-JS}}_{\text{JMI-4}}$	$J^{\text{Ind-JS}}_{\text{CMIM-3}}$	$J^{\text{Ind-JS}}_{\text{CMIM-4}}$
asia	0.828	0.823	0.827	**0.835**
child	0.804	**0.813**	0.739	0.736
hailfinder	0.556	**0.558**	0.486	0.495
alarm	0.704	**0.777**	0.633	0.734
pathfinder	**0.526**	0.497	0.455	0.418
insurance	0.683	0.683	0.654	**0.717**
barley2	0.530	0.510	**0.555**	0.439
andes	0.651	**0.659**	0.600	0.619
win95pts	0.662	**0.684**	0.596	0.563
water	**0.579**	0.570	0.498	0.445
hepar2	**0.658**	0.650	0.644	0.618
Avg ranking	**1.864**	**1.864**	3.182	3.091
CD diagram		Figure 5(b)		

References

Jean Hausser and Korbinian Strimmer. Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research (JMLR)*, 10:1469–1484, 2009.

Olivier Ledoit and Michael Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of empirical finance*, 10(5):603–621, 2003.

James E. Mosimann. On the compound multinomial distribution, the multivariate beta-distribution, and correlations among proportions. *Biometrika*, 49(1/2):65–82, 1962.

Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

Table 4 Comparing the different feature selection methods in terms of their ability to identify the correct features (TPR) for BN with: (a) Sample size = 500 , and (b) Sample size = 2500. The best method (i.e. highest TPR) is highlighted with bold font and at the bottom of the table we present the average ranking score of each method across all datasets.

(a) TPR with 500 sample size

	MIM	MIFS	DISR	ICAP	CIFE	CMI	mRMR	JMI	CMIM	relax-mRMR	JMI-3
asia	0.613	0.683	0.600	0.748	0.763	0.775	0.683	0.755	**0.808**	0.677	0.798
child	0.638	0.562	0.746	0.685	0.648	0.589	0.631	0.675	0.656	0.610	**0.773**
hailfinder	0.464	0.419	**0.503**	0.414	0.413	0.377	0.474	0.468	0.415	0.499	0.497
alarm	0.531	0.482	0.621	0.481	0.609	0.639	0.631	0.632	0.621	0.629	**0.709**
pathfinder	0.460	0.262	0.480	0.352	0.345	0.397	**0.496**	0.494	0.428	0.441	0.450
insurance	0.560	0.488	0.581	0.596	0.609	0.581	0.557	0.628	0.614	0.621	**0.634**
barley2	0.452	0.351	0.412	0.456	0.229	0.245	0.421	0.362	0.373	**0.485**	0.479
andes	0.477	0.422	0.483	0.464	0.536	0.504	0.479	0.562	0.526	0.532	**0.591**
win95pts	0.490	0.361	0.510	0.393	0.483	0.434	0.515	0.584	0.477	0.546	**0.600**
water	0.481	0.382	0.459	0.418	0.389	0.355	0.418	0.445	0.434	0.422	**0.507**
hepar2	0.478	0.381	0.439	0.465	0.452	0.407	0.454	**0.515**	0.485	0.459	0.501
Avg ranking	6.364	9.955	5.545	7.273	7.364	8.000	5.955	3.455	5.364	4.909	**1.818**
CD diagram	Figure 6(a)										

(b) TPR with 2500 sample size

	MIM	MIFS	DISR	ICAP	CIFE	CMI	mRMR	JMI	CMIM	relax-mRMR	JMI-3
asia	0.625	0.700	0.602	0.758	0.802	0.800	0.700	0.775	0.827	0.700	**0.828**
child	0.673	0.567	0.762	0.772	0.772	0.705	0.634	0.733	0.739	0.614	**0.804**
hailfinder	0.503	0.439	0.536	0.454	0.548	0.448	0.500	**0.566**	0.475	0.524	0.556
alarm	0.530	0.480	0.616	0.515	0.646	**0.734**	0.636	0.670	0.633	0.618	0.704
pathfinder	0.485	0.290	0.499	0.393	0.379	0.451	**0.547**	0.547	0.456	0.477	0.526
insurance	0.573	0.500	0.611	0.603	0.621	**0.685**	0.591	0.636	0.652	0.654	0.683
barley2	0.495	0.428	0.526	0.499	0.358	0.336	0.523	0.493	0.508	**0.570**	0.530
andes	0.530	0.494	0.505	0.538	0.605	0.641	0.550	0.622	0.600	0.616	**0.651**
win95pts	0.518	0.409	0.532	0.492	0.594	0.570	0.568	0.627	0.593	0.613	**0.662**
water	0.552	0.380	0.450	0.449	0.440	0.418	0.455	0.565	0.498	0.464	**0.579**
hepar2	0.592	0.447	0.510	0.583	0.630	0.553	0.541	**0.668**	0.644	0.560	0.658
Avg ranking	7.364	10.545	6.909	7.455	5.545	6.182	6.545	3.455	5.000	5.364	**1.636**
CD diagram	Figure 6(b)										

Table 5 Comparing the different FS methods in terms of their misclassification error. In brackets is the ranking score of each method in each dataset, and in the bottom of the table average ranking score of each method across all datasets.

	MIM	MIFS	DISR	ICAP	CIFE	CMI	mRMR	JMI	CMIM	relax-MRMR	JMI-3	CMIM-3	JMI-4	CMIM-4
lungcancer	0.589(9)	0.614(14)	0.589(10)	0.584(5)	0.605(13)	0.570(2)	0.586(7)	0.584(4)	0.585(6)	0.591(11)	0.587(8)	0.571(3)	0.601(12)	0.558(1)
soybeansmall	0.038(6)	0.080(8)	0.037(5)	0.256(14)	0.253(13)	0.103(10)	0.034(1)	0.040(7)	0.081(9)	0.037(4)	0.034(2)	0.105(11)	0.035(3)	0.116(12)
sonar	0.250(8)	0.269(14)	0.260(11)	0.238(3)	0.253(9)	0.266(12)	0.249(7)	0.239(4)	0.238(2)	0.243(5)	0.257(10)	0.233(1)	0.267(13)	0.244(6)
semeion	0.637(14)	0.449(4)	0.490(7)	0.448(3)	0.510(9)	0.459(6)	0.526(11)	0.551(12)	0.457(5)	0.583(13)	0.518(10)	0.445(2)	0.495(8)	0.440(1)
parkinsons	0.120(11)	0.122(13)	0.119(10)	0.112(5)	0.112(6)	0.117(9)	0.117(7)	0.107(2)	0.109(4)	0.117(8)	0.121(12)	0.108(3)	0.125(14)	0.105(1)
ionosphere	0.138(7)	0.138(8)	0.136(5)	0.144(11)	0.149(14)	0.131(2)	0.138(9)	0.138(10)	0.136(6)	0.129(1)	0.136(4)	0.148(12)	0.135(3)	0.148(13)
spect	0.227(8)	0.231(14)	0.229(11)	0.230(13)	0.230(12)	0.228(9)	0.228(10)	0.226(6)	0.227(7)	0.225(3)	0.224(2)	0.225(4)	0.221(1)	0.225(5)
wine	0.071(6)	0.090(12)	0.070(4)	0.101(13)	0.106(14)	0.088(11)	0.069(3)	0.070(5)	0.066(1)	0.068(2)	0.073(7)	0.074(8)	0.077(9)	0.078(10)
breast	0.058(8)	0.061(10)	0.048(2)	0.066(14)	0.066(13)	0.050(4)	0.046(1)	0.051(5)	0.049(3)	0.063(12)	0.059(9)	0.052(6)	0.061(11)	0.053(7)
heart	0.231(6)	0.247(14)	0.234(11)	0.231(8)	0.246(13)	0.243(12)	0.230(4)	0.230(2)	0.228(1)	0.231(7)	0.230(3)	0.231(5)	0.232(9)	0.232(10)
congress	0.066(11)	0.067(12)	0.062(8)	0.070(14)	0.069(13)	0.065(10)	0.060(4)	0.058(3)	0.061(6)	0.061(5)	0.058(1)	0.062(7)	0.058(2)	0.064(9)
musk2	0.095(11)	0.114(14)	0.106(13)	0.093(6)	0.092(5)	0.094(9)	0.104(12)	0.092(4)	0.093(8)	0.089(1)	0.094(10)	0.092(2)	0.093(7)	0.092(3)
splice	0.216(10)	0.237(14)	0.202(6)	0.216(11)	0.224(13)	0.221(12)	0.203(9)	0.202(8)	0.200(1)	0.202(7)	0.201(3)	0.202(5)	0.200(2)	0.201(4)
liver	0.442(10)	0.442(7)	0.442(8)	0.442(10)	0.438(6)	0.445(14)	0.444(13)	0.443(12)	0.442(10)	0.421(1)	0.426(2)	0.431(4)	0.426(3)	0.435(5)
spambase	0.284(13)	0.261(1)	0.264(4)	0.263(2)	0.267(9)	0.267(7)	0.267(6)	0.278(12)	0.267(8)	0.292(14)	0.274(11)	0.264(5)	0.269(10)	0.264(3)
krvskp	0.148(8)	0.166(13)	0.151(11)	0.147(6)	0.126(1)	0.131(2)	0.153(12)	0.142(5)	0.150(10)	0.167(14)	0.138(4)	0.149(9)	0.136(3)	0.147(7)
pima	0.284(3)	0.285(8)	0.286(10)	0.284(2)	0.287(12)	0.288(14)	0.284(4)	0.285(6)	0.284(1)	0.287(13)	0.284(5)	0.285(7)	0.286(9)	0.287(11)
waveform	0.258(1)	0.402(14)	0.231(7)	0.304(12)	0.307(13)	0.280(11)	0.231(6)	0.230(2)	0.232(9)	0.230(4)	0.229(1)	0.230(5)	0.230(3)	0.231(8)
landsat	0.165(14)	0.153(13)	0.146(3)	0.153(12)	0.153(11)	0.147(4)	0.147(5)	0.151(10)	0.147(7)	0.147(6)	0.150(9)	0.145(2)	0.148(8)	0.145(1)
mushroom	0.005(13)	0.006(14)	0.004(11)	0.004(10)	0.003(9)	0.002(8)	0.004(12)	0.002(2)	0.002(6)	0.002(7)	0.002(3)	0.002(1)	0.002(4)	0.002(5)
Avg ranking	9.30	11.05	7.85	8.70	10.40	8.40	7.15	6.05	5.50	6.90	5.80	5.10	6.70	6.10
CD diagram							Figure 7							