Received: 28 December 2020 | Revised: 13 April 2021 | Accepted: 25 April 2021

DOI: 10.1002/sta4.385

ORIGINAL ARTICLE

WILEY

A mutual information criterion with applications to canonical correlation analysis and graphical models

Timothy DelSole[1,2,3] | Michael K. Tippett[4]

[1]Department of Atmospheric, Oceanic, and Earth Sciences, George Mason University, Fairfax, Virginia, 22030, USA

[2]Center for Ocean-Land-Atmosphere Studies, Fairfax, Virginia, 22030, USA

[3]112 Research Hall, Mail Stop 2B3, George Mason University, 4400 University Drive, Fairfax, VA, 22030, USA

[4]Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York, 10027, USA

Corresspondence
Timothy DelSole, Department of Atmospheric, Oceanic, and Earth Sciences, George Mason University, Fairfax, VA, USA.
Email: tdelsole@gmu.edu

Funding information
Climate Program Office, Grant/Award Number: NA14OAR4310160; National Aeronautics and Space Administration, Grant/Award Number: NNX14AM19G; National Science Foundation, Grant/Award Numbers: AGS-1338427, AGS-1822221

This paper derives a criterion for deciding conditional independence that is consistent with small-sample corrections of Akaike's information criterion but is easier to apply to such problems as selecting variables in canonical correlation analysis and selecting graphical models. The criterion reduces to mutual information when the assumed distribution equals the true distribution; hence, it is called mutual information criterion (MIC). Although small-sample Kullback–Leibler criteria for these selection problems have been proposed previously, some of which are not widely known, MIC is strikingly more direct to derive and apply.

KEYWORDS
Akaike's information criterion, CCA, model selection, mutual information

1 | INTRODUCTION

Conditional independence is fundamental to statistical inference (Dawid, 1979). Many tests for conditional independence have been proposed, including tests based on partial correlation, conditional characteristic functions (Su & White, 2007), Hellinger distance (Su & White, 2008), maximal nonlinear conditional correlation (Huang, 2010), and projection-based distance covariance (Fan et al., 2020). These and other criteria are developed from a hypothesis test framework, which has well-known limitations in multiple testing situations (Burnham & Anderson, 2002). An alternative approach to deciding conditional independence is based on Kullback–Leibler (KL) criteria, such as Akaike's information criterion (AIC). AIC is an attractive alternative because it can be applied to multiple testing problems, it does not require specifying an arbitrary significance level, it accounts for out-of-sample variability, and it is derived from a proper score for selecting probability density functions (PDFs) (Akaike, 1973). Nevertheless, applying AIC to decide conditional independence generally requires maximizing a likelihood function subject to the constraint indicated by conditional independence. Such constrained optimization problems can be difficult to solve, which has hindered the application of AIC to such problems.

A clue to a simpler approach comes from regression model selection. In regression model selection, AIC is relatively easy to apply because one simply includes the variables that appear in the regression model and excludes the others. In particular, the relevant AIC does not require solving a constrained optimization problem. For selection problems that cannot be reduced to regression model selection, the question arises as to whether there exists a criterion similar to AIC that does not require solving constrained maximum likelihood problems, and yet can be

evaluated by excluding the variables that are conditionally independent of the retained variables. The purpose of this paper is to derive such a criterion.

We begin by seeking a criterion whose differences equal the differences in KL divergences in the case of selecting explanatory variables of a regression model. This ensures that the criterion recovers regression model selection. Then, we add one more condition, namely, that the criterion should be symmetric, in the sense that the criterion does not depend on which variables are labelled response and explanatory. Remarkably, only one quantity satisfies these conditions. This quantity reduces to mutual information when the model PDF equals the true PDF. Accordingly, we call this quantity mutual information criterion ($\mathbb{MIC}$). This paper demonstrates that $\mathbb{MIC}$ is our desired criterion.

Sample estimates of $\mathbb{MIC}$ can be derived based on AIC. Naturally, the resulting estimates share the same limitations as AIC. One well-known limitation of AIC is that it tends to select overfitted models. A standard fix to this problem is to use a small-sample corrected version called AICc (Hurvich & Tsai, 1989). Unfortunately, AICc implicitly assumes that the explanatory variables are the same between training and verification samples (DelSole & Tippett, 2021; Rosset & Tibshirani, 2020; Tian et al., 2020). We show that this assumption implies that AICc is not guaranteed to make consistent decisions about conditional independence. Therefore, AICc is not appropriate for estimating $\mathbb{MIC}$. The appropriate small-sample correction to AIC that accounts for independent training and verification samples has been derived recently by DelSole and Tippett (2021) and Tian et al. (2020) (here called AICr). This criterion is used to derive an estimate of $\mathbb{MIC}$, called MIC. Because MIC is based on the newly derived AICr rather than AIC or AICc, it improves upon previous criteria even for the extensively studied case of regression model selection.

The problem of selecting both response and explanatory variables is more formidable. However, MIC provides a very reasonable small-sample criterion for selecting explanatory and response variables and is well suited for selecting variables for canonical correlation analysis (CCA). Another application of MIC is to select graphical models. Graphical models provide a visual summary of various conditional independencies among variables. Conditional independence implies that an associated conditional mutual information vanishes. We derive an analogous criterion called conditional MIC that provides a small-sample criterion for selecting graphical models.

In a series of papers, Yasunori Fujikoshi derived small-sample criteria for many of the above selection problems by explicitly maximizing the likelihood function under the appropriate hypothesis of conditional independence (Fujikoshi, 1982, 1985; Fujikoshi et al., 2010). We show that differences in MIC are equivalent to each of these criteria derived by Fujikoshi (after accounting for slight differences in formulation). Despite these earlier derivations, the derivation presented here is of considerable value because of its greater simplicity compared to previous derivations. The basis of this simplification is that KL divergences satisfy certain identities called chain rules. These chain rules can be used to convert certain constrained maximum likelihood problems into unconstrained problems. As a result, small-sample criteria for conditional independence can be derived from these chain rules, thereby avoiding direct maximization of the likelihood function, which often requires intricate matrix manipulation.

2 | DERIVATION OF THE NEW CRITERION

Let $\mathbf{x}$ and $\mathbf{y}$ be random vectors with a joint PDF $p(\mathbf{x}, \mathbf{y})$. In practice, the true PDF is unknown. Our goal is to identify an approximate PDF by deciding if the PDF has *structure* and then to estimate the PDF under this constraint. Let $q(\mathbf{x}, \mathbf{y})$ denote a candidate PDF without structure, and let $q_1(\mathbf{x}, \mathbf{y})$, $q_2(\mathbf{x}, \mathbf{y})$, … denote candidate PDFs with different structures. Our criterion for choosing a particular structure is that it minimizes the KL divergence or equivalently minimizes

$$\mathbb{H}_i(XY) = -2\mathbb{E}_{XY}[\log q_i(\mathbf{x}, \mathbf{y})], \tag{1}$$

where $\mathbb{E}_{XY}[\cdot]$ denotes the expectation with respect to $p(\mathbf{x}, \mathbf{y})$. $\mathbb{H}_i(XY)$ is called the *cross entropy* between p and q_i (ignoring an irrelevant factor of 2). $\mathbb{H}(XY)$ (with no subscript) denotes the cross entropy between p and q. When $p = q$, cross entropy equals (twice) the entropy of $p(\mathbf{x}, \mathbf{y})$.

The criterion for selecting structure is well developed in the special case of selecting regression models. To be precise, the selection of regression models will be called *X*-selection and defined as follows.

Definition 1 ***X*-selection.** A regression model (also called a prediction model) is effectively a conditional PDF $q_i(\mathbf{y}|\mathbf{x})$, where the first and second variables are called response and explanatory, respectively. The prediction model is related to the joint PDF as

$$q_i(\mathbf{x}, \mathbf{y}) = q_i(\mathbf{y}|\mathbf{x})q_i(\mathbf{x}). \tag{2}$$

The *X*-selection problem is to select one prediction model from a set of candidate models $q_1(\mathbf{y}|\mathbf{x}_1)$, $q_2(\mathbf{y}|\mathbf{x}_2)$, …. The candidate PDFs are restricted to ones in which the prediction models differ in their explanatory variables $\mathbf{x}_1$, $\mathbf{x}_2$, …, each of which is a subset of $\mathbf{x}$, but have the same response variable $\mathbf{y}$. It is assumed that each prediction model equals the unconstrained PDF conditioned on the appropriate subset of explanatory variables:

$$q_i(\mathbf{y}|\mathbf{x}) = q_i(\mathbf{y}|\mathbf{x}_i) = q(\mathbf{y}|\mathbf{x}_i). \tag{3}$$

The first equality states that certain X variables may be omitted from $q_i(\mathbf{y}|\mathbf{x})$ without changing the prediction, and the second equality states that the resulting prediction model equals the unconstrained PDF $q(\mathbf{y}|\mathbf{x}_i)$. Aside from this, no further structure is imposed. In particular, no structure is imposed on $q_i(\mathbf{x})$:

$$q_i(\mathbf{x}) = q(\mathbf{x}) \text{ for all } i. \tag{4}$$

It follows from (2)–(4) that

$$q_i(\mathbf{x}, \mathbf{y}) = q(\mathbf{x})q(\mathbf{y}|\mathbf{x}_i). \tag{5}$$

This identity shows that the joint PDF can be written as a product of PDFs where structure is imposed by omitting X variables in the conditional PDF. Note that the joint PDF $q_i(\mathbf{x}, \mathbf{y})$ depends on the full $\mathbf{x}$ even when the prediction model $q(\mathbf{y}|\mathbf{x}_i)$ depends only on a proper subset of $\mathbf{x}$. Variables that can be omitted from conditionals are said to be *redundant*.

Lemma 1 The chain rule lemma. It follows from (1) and (2) that cross entropy satisfies the chain rule

$$\mathbb{H}_i(XY) = \mathbb{H}_i(X) + \mathbb{H}_i(Y|X), \tag{6}$$

where the cross entropy for prediction models is $\mathbb{H}_i(Y|X) = -2\mathbb{E}_{XY}[\log q_i(\mathbf{y}|\mathbf{x}_i)]$. Under X-selection,

$$\mathbb{H}_i(XY) = \mathbb{H}(X) + \mathbb{H}(Y|X_i), \tag{7}$$

where $\mathbb{H}_i(X) = \mathbb{H}(X)$ follows from (4), and $\mathbb{H}_i(Y|X) = \mathbb{H}(Y|X_i)$ follows from (3).

Lemma 1 implies that under X-selection,

$$\mathbb{H}_i(XY) - \mathbb{H}_j(XY) = \mathbb{H}(Y|X_i) - \mathbb{H}(Y|X_j). \tag{8}$$

Because only differences in cross entropy affect selection, this identity shows that, under X-selection, selecting prediction models based on $\mathbb{H}(Y|X_i)$ is equivalent to selecting structured PDFs based on $\mathbb{H}_i(XY)$. Importantly, the left-hand side of (8) involves structured PDFs while the right hand side involves only unstructured PDFs. This fact will become important later when we derive estimates of cross entropy—the left-hand side will require solving constrained maximum likelihood problems, whereas the right-hand side will not.

Not all selection problems can be reduced to X-selection. For instance, in CCA, both X and Y variables are selected. We call this *simultaneous selection*. $\mathbb{H}(Y|X)$ is not a meaningful criterion for simultaneous selection because Y differs between models. For instance, $\mathbb{H}(Y|X)$ is a proxy for prediction error, and comparing prediction errors of different quantities with different units is not meaningful. In such cases, the natural approach is to define the structure in $q_i(\mathbf{x}, \mathbf{y})$ associated with the selection problem and then compute the corresponding cross entropy $\mathbb{H}_i(XY)$. However, this approach inevitably leads to solving a constrained maximum likelihood problem, which can be difficult. We seek an alternative approach that avoids solving a constrained maximum likelihood problem, similar to the way regression model selection avoids this problem. More precisely, we seek a criterion that can be computed by omitting redundant X and Y variables from the calculation, just as $\mathbb{H}(Y|X)$ can be computed by omitting redundant X variables from the prediction model. Let this new criterion be denoted $\mathbb{MIC}(X;Y)$, where explanatory and response variables are separated by a semicolon. The first natural requirement is that it should be consistent with cross entropy for X-selection.

Definition 2. $\mathbb{MIC}(X;Y)$ is said to be consistent with cross entropy for X-selection if for all $q(\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2)$ and $p(\mathbf{y}, \mathbf{x}_1, \mathbf{x}_2)$,

$$\mathbb{MIC}(X_1;Y) - \mathbb{MIC}(X_2;Y) = \mathbb{H}(Y|X_1) - \mathbb{H}(Y|X_2). \tag{9}$$

A second requirement is that it should be suitable for simultaneous selection, particularly for selecting variables in CCA. Importantly, CCA does not distinguish response and explanatory variables. Therefore, we seek a criterion that satisfies the following property.

Definition 3 Symmetric. A selection criterion is said to be *symmetric* if it does not depend on which variables are labelled response and explanatory.

Clearly, $\mathbb{H}(Y|X)$ is not symmetric, since $\mathbb{H}(Y|X) = -2\mathbb{E}_{XY}[\log q_{Y|X}(\mathbf{y}|\mathbf{x})] \neq -2\mathbb{E}_{XY}[\log q_{X|Y}(\mathbf{x}|\mathbf{y})] = \mathbb{H}(X|Y)$. On the other hand, $\mathbb{H}(XY)$ is symmetric, but it is not consistent with cross entropy since $\mathbb{H}(X_1Y) - \mathbb{H}(X_2Y) = \mathbb{H}(Y|X_1) - \mathbb{H}(Y|X_2) + \mathbb{H}(X_1) - \mathbb{H}(X_2)$. In general, $\mathbb{H}(X_1) - \mathbb{H}(X_2) \neq 0$. The criterion that is both symmetric and consistent with cross entropy is given in the following proposition.

Proposition 1. To within an additive constant, the only criterion that is both symmetric and consistent with cross entropy for X-selection is

$$\mathbb{MIC}(X;Y) = \mathbb{H}(XY) - \mathbb{H}(Y) - \mathbb{H}(X) \tag{10}$$

$$= \mathbb{H}(Y|X) - \mathbb{H}(Y) \tag{11}$$

$$= \mathbb{H}(X|Y) - \mathbb{H}(X). \tag{12}$$

Proof. Let $\mathbb{H}(Y|X_1X_2)$ and $\mathbb{H}(Y|X_1)$ denote cross entropies for $q(\mathbf{y}|\mathbf{x}_1, \mathbf{x}_2)$ and $q(\mathbf{y}|\mathbf{x}_1)$, respectively. By assumption, $\mathbb{MIC}(X;Y)$ is consistent with cross entropy for X-selection; hence,

$$\mathbb{MIC}(X_1X_2;Y) - \mathbb{MIC}(X_1;Y) = \mathbb{H}(Y|X_1X_2) - \mathbb{H}(Y|X_1). \tag{13}$$

Rearranging this equation gives

$$\mathbb{MIC}(X_1X_2;Y) - \mathbb{H}(Y|X_1X_2) = \mathbb{MIC}(X_1;Y) - \mathbb{H}(Y|X_1). \tag{14}$$

The absence of X_2 on the right implies that the right-hand side is a functional of the distribution of $\mathbf{x}_1$ and $\mathbf{y}$ only. It follows that the left-hand side has this same dependence. Repeating the above argument but with the roles of $\mathbf{x}_1$ and $\mathbf{x}_2$ swapped leads to the conclusion that the left-hand side is a functional of the joint distribution only of $\mathbf{x}_2$ and $\mathbf{y}$. These two properties hold for arbitrary $p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{y})$ only if the left-hand side is a functional of the distribution of $\mathbf{y}$ only. That is,

$$\mathbb{MIC}(X;Y) - \mathbb{H}(Y|X) = f(Y), \tag{15}$$

where $f(Y)$ is some functional of $q(\mathbf{y})$ and $p(\mathbf{y})$. Similar arguments, but swapping the roles of X and Y, give

$$\mathbb{MIC}(Y;X) - \mathbb{H}(X|Y) = g(X), \tag{16}$$

where $g(X)$ is some functional of $q(X)$ and $p(X)$. By assumption, $\mathbb{MIC}$ is symmetric; hence, $\mathbb{MIC}(X;Y) = \mathbb{MIC}(Y;X)$. Therefore, $\mathbb{MIC}$ may be eliminated from (15) and (16) to give

$$\mathbb{H}(Y|X) + f(Y) = \mathbb{H}(X|Y) + g(X). \tag{17}$$

Substituting the chain rule (6) into (17) and rearranging terms gives

$$\mathbb{H}(X) + g(X) = \mathbb{H}(Y) + f(Y). \tag{18}$$

The left-hand side does not depend on the distribution of Y, and the right-hand side does not depend on the distribution of X. The only way that this identity can hold for arbitrary distributions is that the two sides must equal a constant. Therefore,

$$\mathbb{H}(Y) + f(Y) = \alpha,$$

where α is a constant. Since only differences in $\mathbb{MIC}$ are important, the constant α may be set to zero without loss of generality. Solving for $f(Y)$ and substituting into (15) determines MIC uniquely and yields (11). Equations (10) and (12) follow from (11) by the chain rule (6).

To our knowledge, $\mathbb{MIC}$ has not appeared in the literature. If $p(\mathbf{x},\mathbf{y}) = q(\mathbf{x},\mathbf{y})$, then $\mathbb{MIC}(X;Y) = -2\mathbb{M}(X;Y)$, where

$$\mathbb{M}(X;Y) = \mathbb{E}_{XY}\left[\log\frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}\right]$$

is the *mutual information* between $\mathbf{x}$ and $\mathbf{y}$. Just as $\mathbb{H}$ is cross entropy (times 2), $\mathbb{MIC}$ may be called *cross mutual information* (times -2). Anticipating its application to variable selection, we call $\mathbb{MIC}$ mutual information criterion. The explicit dependence of $\mathbb{MIC}(X;Y)$ on the model PDF is

$$\mathbb{MIC}(X;Y) = -2\mathbb{E}_{XY}\left[\log\frac{q(\mathbf{x},\mathbf{y})}{q(\mathbf{x})q(\mathbf{y})}\right] = -2\mathbb{E}_{XY}\left[\log\frac{q(\mathbf{y}|\mathbf{x})}{q(\mathbf{y})}\right]. \tag{19}$$

Conditional $\mathbb{MIC}$ can be defined analogously to conditional mutual information:

$$\mathbb{MIC}(X;Y|Z) = \mathbb{H}(Y|X,Z) - \mathbb{H}(Y|Z) = -2\mathbb{E}_{XYZ}\left[\log\frac{q(\mathbf{x},\mathbf{y}|\mathbf{z})}{q(\mathbf{x}|\mathbf{z})q(\mathbf{y}|\mathbf{z})}\right]. \tag{20}$$

$\mathbb{MIC}$ satisfies chain rules analogous to mutual information; for example,

$$\mathbb{MIC}(XZ;Y) = \mathbb{MIC}(X;Y) + \mathbb{MIC}(Z;Y|X). \tag{21}$$

3 | VARIABLE SELECTION AND CONDITIONAL INDEPENDENCE

Although MIC is consistent with cross entropy for X-selection, this does not guarantee that it is a sensible criterion for simultaneous selection. To show the latter, we first clarify the structure associated with X-selection.

Definition 4. X variables will be partitioned as $\mathbf{x} = (\mathbf{x}_K^T\mathbf{x}_R^T)^T$, where $\mathbf{x}_K$ denotes the M_K variables to keep and $\mathbf{x}_R$ denotes M_R variables either to remove or retain. Similarly, Y variables will be partitioned as $\mathbf{y} = (\mathbf{y}_K^T,\mathbf{y}_R^T)^T$, where $\mathbf{y}_K$ and $\mathbf{y}_R$ have dimensions P_K and P_R.

Under X-selection, the decision to remove $\mathbf{x}_R$ from the prediction model depends on the cross entropies of $p(\mathbf{y}|\mathbf{x}_K, \mathbf{x}_R)$ and $p(\mathbf{y}|\mathbf{x}_K)$. The structure relevant to this problem is (3), which is expressed below in the notation of Definition 4:

$$q_\omega(\mathbf{y}|\mathbf{x}_K,\mathbf{x}_R) = q_\omega(\mathbf{y}|\mathbf{x}_K) = q(\mathbf{y}|\mathbf{x}_K), \tag{22}$$

where ω denotes the appropriate structural constraint on the PDF. Under (22), $\mathbb{H}_\omega(Y|X_KX_R) = \mathbb{H}(Y|X_K)$, and therefore,

$$\mathbb{H}(Y|X_KX_R) - \mathbb{H}_\omega(Y|X_KX_R) = \mathbb{H}(Y|X_KX_R) - \mathbb{H}(Y|X_K),$$

which shows that using cross entropy to decide to remove $\mathbf{x}_R$ is indistinguishable from deciding that the model PDF satisfies (22). The first equality in (22) asserts that, under the model PDF, $\mathbf{y}$ and $\mathbf{x}_R$ are *conditionally independent* given $\mathbf{x}_K$. We denote this condition as

$$Y \perp \mathrm{X_R}|\mathrm{X_K}. \tag{23}$$

Conditional independence defines a particular structure on a PDF. Importantly, conditional independence can be expressed through $q(\cdot)$ in different ways. For instance, by repeated application of the probability law (2), the structure (22) can be expressed equivalently as

$$q_\omega(\mathbf{y}|\mathbf{x}_R,\mathbf{x}_K) = q(\mathbf{y}|\mathbf{x}_K), \tag{24}$$

$$q_\omega(\mathbf{x}_R|\mathbf{y},\mathbf{x}_K) = q(\mathbf{x}_R|\mathbf{x}_K), \tag{25}$$

$$q_\omega(\mathbf{y},\mathbf{x}_R|\mathbf{x}_K) = q(\mathbf{y}|\mathbf{x}_K)q(\mathbf{x}_R|\mathbf{x}_K), \tag{26}$$

$$q_\omega(\mathbf{y},\mathbf{x}_R,\mathbf{x}_K) = q(\mathbf{y},\mathbf{x}_K)q(\mathbf{x}_R,\mathbf{x}_K)/q(\mathbf{x}_K). \tag{27}$$

These expressions are equivalent statements that the model PDF satisfies $Y \perp X_R \mid X_K$. This equivalence allows us to prove the following.

Proposition 2. Under conditional independence $\omega : Y \perp X_R \mid X_K$,

$$\mathbb{MIC}(X_K X_R; Y) - \mathbb{MIC}(X_K; Y) = \mathbb{H}(X_K X_R Y) - \mathbb{H}_\omega(X_K X_R Y). \tag{28}$$

where $\mathbb{H}_\omega(X_K X_R Y)$ is the cross-entropy of $q_\omega(\mathbf{x}_K, \mathbf{x}_R, \mathbf{y})$ defined in (27).

Proof. Computing the cross entropy of (27) yields $\mathbb{H}_\omega(YX_RX_K) = \mathbb{H}(YX_K) + \mathbb{H}(X_RX_K) - \mathbb{H}(X_K)$, and therefore,

$$\begin{aligned}\mathbb{H}(X_K X_R Y) - \mathbb{H}_\omega(X_K X_R Y) &= \mathbb{H}(X_K X_R Y) - (\mathbb{H}(YX_K) + \mathbb{H}(X_R X_K) - \mathbb{H}(X_K)) \\ &= (\mathbb{H}(X_K X_R Y) - \mathbb{H}(X_R X_K) - \mathbb{H}(Y)) - (\mathbb{H}(YX_K) - \mathbb{H}(X_K) - \mathbb{H}(Y)) \\ &= \mathbb{MIC}(X_K X_R; Y) - \mathbb{MIC}(X_K; Y),\end{aligned}$$

which proves the proposition.

Proposition 2 shows that $\mathbb{MIC}$ is consistent with cross entropy for deciding conditional independence (23). By analogy, we anticipate that simultaneous selection corresponds to selecting some form of conditional independence. To define this form, note that simultaneous selection asks whether ($\mathbf{x}_R$; $\mathbf{y}_R$) should be included with ($\mathbf{x}_K$; $\mathbf{y}_K$). By analogy with X-selection, the criterion for simultaneous selection should be based on comparing MIC with and without the potentially redundant variables ($\mathbf{x}_R$; $\mathbf{y}_R$), that is, based on comparing $\mathbb{MIC}(X_K X_R; Y_K Y_R)$ to $\mathbb{MIC}(X_K; Y_K)$. The structure required for this difference in $\mathbb{MIC}$ to equal the difference in cross entropies between unstructured and structured PDFs is given next.

Proposition 3. The criterion

$$\mathbb{MIC}(X_K X_R; Y_K Y_R) - \mathbb{MIC}(X_K; Y_K) = \mathbb{H}(X_K X_R Y_K Y_R) - \mathbb{H}_\psi(X_K X_R Y_K Y_R) \tag{29}$$

holds if and only if the constraint ψ is

$$\psi : Y_K \perp \mathrm{X_R}|\mathrm{X_K} \ \text{ and } \ \mathrm{Y_R} \perp \mathrm{X_K X_R}|\mathrm{Y_K}. \tag{30}$$

For clarity, we note that (30) can be expressed in other equivalent forms using logical equivalences (an example is 73 below; see Dawid, 1979).

Proof. Expanding $\mathbb{MIC}$ using definition (10) and rearranging terms gives

$$\begin{aligned}\mathbb{MIC}(X_K X_R; Y_K Y_R) - \mathbb{MIC}(X_K; Y_K) &= (\mathbb{H}(X_K X_R Y_K Y_R) - \mathbb{H}(Y_K Y_R) - \mathbb{H}(X_K X_R)) - (\mathbb{H}(X_K Y_K) - \mathbb{H}(X_K) - \mathbb{H}(Y_K)) \\ &= \mathbb{H}(X_K X_R Y_K Y_R) - (\mathbb{H}(Y_R|Y_K) + \mathbb{H}(Y_K|X_K) + \mathbb{H}(X_K X_R)).\end{aligned} \tag{31}$$

Comparison with (29) implies

$$\mathbb{H}_\psi(X_K X_R Y_K Y_R) = \mathbb{H}(Y_R|Y_K) + \mathbb{H}(Y_K|X_K) + \mathbb{H}(X_K X_R), \tag{32}$$

or equivalently

$$\begin{aligned} 0 &= \mathbb{H}_\psi(X_K X_R Y_K Y_R) - (\mathbb{H}(Y_R|Y_K) + \mathbb{H}(Y_K|X_K) + \mathbb{H}(X_K X_R)) \\ &= \mathbb{E}_{X_K X_R Y_K Y_R}\left[\log\left(\frac{q_\psi(\mathbf{x}_K,\mathbf{x}_R,\mathbf{y}_K,\mathbf{y}_R)}{q(\mathbf{y}_R|\mathbf{y}_K)q(\mathbf{y}_K|\mathbf{x}_K)q(\mathbf{x}_K,\mathbf{x}_R)}\right)\right] \\ &= \mathbb{E}_{X_K X_R Y_K Y_R}\left[\log\left[\left(\frac{q_\psi(\mathbf{y}_R|\mathbf{x}_K,\mathbf{x}_R,\mathbf{y}_K)}{q(\mathbf{y}_R|\mathbf{y}_K)}\right)\left(\frac{q_\psi(\mathbf{y}_K|\mathbf{x}_K,\mathbf{x}_R)}{q(\mathbf{y}_K|\mathbf{x}_K)}\right)\left(\frac{q_\psi(\mathbf{x}_K,\mathbf{x}_R)}{q(\mathbf{x}_K,\mathbf{x}_R)}\right)\right]\right]. \end{aligned}$$

For the expectation to vanish for any true PDF $p(\mathbf{x}_K, \mathbf{x}_R, \mathbf{y}_K, \mathbf{y}_R)$, the argument of the log must equal one. The first parenthesis is the only term that depends on $\mathbf{y}_R$. By familiar arguments in separation of variables, this term must equal a constant and that constant must be one to ensure that the model PDFs integrate to one. Under this result, the second parenthesis is the only term that depends on $\mathbf{y}_K$; hence, by similar arguments, it too must equal one. Given these two results, the last term in parenthesis must equal one, implying that ψ does not impose structure on $q(\mathbf{x})$. It follows that

$$q_\psi(\mathbf{y}_R|\mathbf{y}_K,\mathbf{x}_K,\mathbf{x}_R) = q(\mathbf{y}_R|\mathbf{y}_K) \Leftrightarrow Y_R \perp \mathrm{X_K X_R}|\mathrm{Y_K}, \tag{33}$$

$$q_\psi(\mathbf{y}_K|\mathbf{x}_K,\mathbf{x}_R) = q(\mathbf{y}_K|\mathbf{x}_K) \Leftrightarrow Y_K \perp \mathrm{X_R}|\mathrm{X_K}, \tag{34}$$

which are the constraints in (30). The corresponding constrained joint PDF is

$$q_\psi(\mathbf{x}_K,\mathbf{x}_R,\mathbf{y}_K,\mathbf{y}_R) = q(\mathbf{y}_R|\mathbf{y}_K)q(\mathbf{y}_K|\mathbf{x}_K)q(\mathbf{x}_K,\mathbf{x}_R). \tag{35}$$

This proves the "only if" part. To prove the "if" part, note that ψ in (30) implies (35), which implies (32), which if substituted in (29) yields (31).

To clarify the reasonableness of (23) and (30), the following proposition describes their consequences in terms of CCA.

Proposition 4 Adding redundant variables to x_K and y_K does not alter the canonical correlations.. Consider CCA of $\mathbf{x}$ and $\mathbf{y}$, which yields a projection vector pair $\mathbf{u}$ and $\mathbf{v}$ such that the correlation between $\mathbf{u}^T\mathbf{x}$ and $\mathbf{v}^T\mathbf{y}$ equals the canonical correlation. Following Definition 4, partition $\mathbf{u} = (\mathbf{u}_K^T \mathbf{u}_R^T)^T$ and $\mathbf{v} = (\mathbf{v}_K^T \mathbf{v}_R^T)^T$. If (23) is true, then $\mathbf{u}_R = \mathbf{0}$ for all canonical correlations. If (30) is true, then $\mathbf{u}_R = \mathbf{0}$ and $\mathbf{v}_R = \mathbf{0}$ for all canonical correlations. In either case, the canonical correlations for $(\mathbf{x}_K;\ \mathbf{y}_K)$ are identical to those of $(\mathbf{x};\mathbf{y})$.

Proof. The constraint ψ in (30) can be written in terms of $q_\psi(\cdot)$ as in (33) and (34), which in turn can be written, respectively, as

$$q_\psi(\mathbf{y}_R,\mathbf{x}|\mathbf{y}_K) = q_\psi(\mathbf{y}_R|\mathbf{y}_K)q_\psi(\mathbf{x}|\mathbf{y}_K) \ \text{ and } \ q_\psi(\mathbf{y}_K,\mathbf{x}_R|\mathbf{x}_K) = q_\psi(\mathbf{y}_K|\mathbf{x}_K)q_\psi(\mathbf{x}_R|\mathbf{x}_K). \tag{36}$$

Let $\mathrm{cov}_\psi[\mathbf{y}_R,\mathbf{x}|\mathbf{y}_K]$ denote the conditional covariance matrix between $\mathbf{y}_R$ and $\mathbf{x}$ given $\mathbf{y}_K$ under model PDF $q_\psi(\mathbf{x}_K,\mathbf{x}_R,\mathbf{y}_K,\mathbf{y}_R)$. Then (36) implies

$$\mathrm{cov}_\psi[\mathbf{y}_R,\mathbf{x}|\mathbf{y}_K] = 0 \ \text{ and } \ \mathrm{cov}_\psi[\mathbf{y}_K,\mathbf{x}_R|\mathbf{x}_K] = 0. \tag{37}$$

Under covariance constraints (37), Fujikoshi (1982) showed that $\mathbf{u}_R = \mathbf{0}$ and $\mathbf{v}_R = \mathbf{0}$. Under the second identity in (37), Fujikoshi et al. (2010) showed that $\mathbf{u}_R = \mathbf{0}$. In both cases, the canonical correlations for $(\mathbf{x}_K;\ \mathbf{y}_K)$ are identical to those of $(\mathbf{x};\mathbf{y})$. This completes the proof.

4 | SAMPLE CRITERION FOR NORMAL DISTRIBUTIONS

The above considerations have ignored the fact that model PDFs generally involve parameters that are unknown and must be estimated from finite samples. This estimation can lead to overfitting and must be taken into account. Let $q(\mathbf{Y}|\mathbf{X};\boldsymbol{\theta}_{Y|X})$ denote the PDF model for predicting $\mathbf{Y}$ given $\mathbf{X}$ with parameters $\boldsymbol{\theta}_{Y|X}$. We follow Akaike (1973) by using maximum likelihood estimates (MLEs) for the parameters. Accordingly, let $\hat{\boldsymbol{\theta}}_{Y|X}$ denote the MLE of $\boldsymbol{\theta}_{Y|X}$ derived from the sample $(\hat{\mathbf{X}},\hat{\mathbf{Y}})$. A fundamental principle in model selection is to judge model performance based on how well the model predicts an *independent* sample $(\mathbf{X}_0, \mathbf{Y}_0)$. Following Akaike (1973), we average the cross entropy for $q(\mathbf{Y}_0|\mathbf{X}_0;\hat{\boldsymbol{\theta}}_{Y|X})$ over $(\hat{\mathbf{X}},\hat{\mathbf{Y}})$ and $(\mathbf{X}_0, \mathbf{Y}_0)$, which have identical distributions but are independent of each other. The result is $\mathbb{IC}$:

$$\mathbb{IC}(Y|X) = -2\mathbb{E}_{\hat{X}\hat{Y}}\left[\mathbb{E}_{X_0Y_0}\left[\log q(\mathbf{Y}_0|\mathbf{X}_0;\hat{\boldsymbol{\theta}}_{Y|X})\right]\right]. \tag{38}$$

Under normality, the PDF model satisfies $q(\mathbf{X},\mathbf{Y};\hat{\boldsymbol{\theta}}_{XY}) = q(\mathbf{X};\hat{\boldsymbol{\theta}}_X)q(\mathbf{Y}|\mathbf{X};\hat{\boldsymbol{\theta}}_{Y|X})$, where $\hat{\boldsymbol{\theta}}_{XY},\hat{\boldsymbol{\theta}}_X,\hat{\boldsymbol{\theta}}_{Y|X}$ are MLEs of the parameters in the respective PDF models (this identity does not hold in general; Barndorff-Nielsen, 1976). As a result of this identity, $\mathbb{IC}$ satisfies the chain rule

$$\mathbb{IC}(XY) = \mathbb{IC}(X) + \mathbb{IC}(Y|X), \tag{39}$$

where $\mathbb{IC}(XY) = -2\mathbb{E}_{\hat{X}\hat{Y}}[\mathbb{E}_{X_0Y_0}[\log q(\mathbf{X}_0,\mathbf{Y}_0;\hat{\boldsymbol{\theta}}_{XY})]]$ and $\mathbb{IC}(X) = -2\mathbb{E}_{\hat{X}}[\mathbb{E}_{X_0}[\log q(\mathbf{X}_0;\hat{\boldsymbol{\theta}}_X)]]$. By analogy, we define the following:

Proposition 5. The Akaike-type extension of $\mathbb{MIC}$ is defined as

$$\mathbb{MIC}\text{a}(X;Y) = \mathbb{IC}(Y|X) - \mathbb{IC}(Y) = -2\mathbb{E}_{\hat{X}\hat{Y}}\left[\mathbb{E}_{X_0Y_0}\left[\log\frac{q(\mathbf{Y}_0|\mathbf{X}_0;\hat{\boldsymbol{\theta}}_{Y|X})}{q(\mathbf{Y}_0;\hat{\boldsymbol{\theta}}_Y)}\right]\right]. \tag{40}$$

To within an additive constant, the only criterion that is both symmetric and whose differences equal the corresponding differences in $\mathbb{IC}$ is $\mathbb{MIC}$a.

Proof. In the proof for Proposition 1, replace $\mathbb{H}$ everywhere by $\mathbb{IC}$. Then, the proof follows the same steps. In particular, the analogous expression for (14) has the right-hand side $\mathbb{MIC}\text{a}(X_K;Y) - \mathbb{IC}(Y|X_K)$, which still is a functional of the distribution of $\mathbf{x}_K$ and $\mathbf{y}$ only, because $q(\mathbf{y}|\mathbf{x}_K;\hat{\boldsymbol{\theta}}_{Y|X_K})$ does not depend on $\mathbf{x}_R$. Also, $\mathbb{IC}$ satisfies the chain rule (39), so the step from (17) to (18) is essentially the same as for $\mathbb{H}$.

Proposition 5 implies that estimates of $\mathbb{MIC}$a follow from estimates of $\mathbb{IC}$, and so we consider in some detail unbiased and consistent estimation of $\mathbb{IC}$. For normal distributions, such estimates can be derived from the model,

$$\mathbf{Y} = \mathbf{XB} + \mathbf{j}\boldsymbol{\mu}_Y^T + \mathbf{E}_Y, \tag{41}$$

where $\mathbf{Y}$ and $\mathbf{X}$ are identified as response and explanatory variables, respectively, $\mathbf{B}$ and $\boldsymbol{\mu}_Y$ contain regression coefficients, $\mathbf{j}$ is a vector of ones to account for the intercept, and $\mathbf{E}_Y$ is a random matrix. Each row of $\mathbf{E}_Y$ is independently distributed as a multivariate normal with zero mean and covariance matrix $\boldsymbol{\Sigma}_{Y|X}$. The dimensions are

$$\mathbf{Y} \in \mathbb{R}^{N\times P},\ \mathbf{X} \in \mathbb{R}^{N\times M_K},\ \mathbf{B} \in \mathbb{R}^{M_K\times P},\ \mathbf{j} \in \mathbb{R}^N,\ \boldsymbol{\mu}_Y \in \mathbb{R}^P,\ \mathbf{E}_Y \in \mathbb{R}^{N\times P}.$$

The total number of predictors including the intercept is $M = M_K + 1$. Sugiura (1978) and Hurvich and Tsai (1989) showed that if the candidate model (41) includes the true model (and if other restrictions discussed below hold), then an unbiased estimate of (38) is

$$\text{AICc}(Y|X) = N\log|\hat{\boldsymbol{\Sigma}}_{Y|X}| + NP\log(2\pi) + NP + N\frac{2MP + P(P+1)}{N-M-P-1}, \tag{42}$$

where $\hat{\boldsymbol{\Sigma}}_{Y|X}$ is the MLE of $\boldsymbol{\Sigma}_{Y|X}$ derived from $(\hat{\mathbf{X}},\hat{\mathbf{Y}})$. Estimates of $\mathbb{IC}(Y)$, $\mathbb{IC}(X)$, $\mathbb{IC}(XY)$ may be derived by applying (42) to the models

$$\mathbf{Y} = \mathbf{j}\boldsymbol{\mu}_Y^T + \mathbf{E}_Y,\ \ \mathbf{X} = \mathbf{j}\boldsymbol{\mu}_X^T + \mathbf{E}_X,\ \ [\mathbf{XY}] = \mathbf{j}[\boldsymbol{\mu}_X^T\boldsymbol{\mu}_Y^T] + \mathbf{E}_{XY}. \tag{43}$$

Let $\hat{\boldsymbol{\Sigma}}_{YY},\hat{\boldsymbol{\Sigma}}_{XX},\hat{\boldsymbol{\Sigma}}_{(XY)}$ be the MLEs of the covariance matrices of $\mathbf{E}_Y,\mathbf{E}_X,\mathbf{E}_{XY}$, respectively. These matrices are related through standard identities

$$\hat{\boldsymbol{\Sigma}}_{Y|X} = \hat{\boldsymbol{\Sigma}}_{YY} - \hat{\boldsymbol{\Sigma}}_{YX}\hat{\boldsymbol{\Sigma}}_{XX}^{-1}\hat{\boldsymbol{\Sigma}}_{XY}, \tag{44}$$

$$|\hat{\boldsymbol{\Sigma}}_{(XY)}| = |\hat{\boldsymbol{\Sigma}}_{XX}||\hat{\boldsymbol{\Sigma}}_{Y|X}|. \tag{45}$$

Evaluating AICc for each model in (43), noting that each model has only $M = 1$ explanatory variable (i.e., the intercept), and using the identity $NP + N(2MP + P(P+1))/(N-M-P-1) = PN(N+M)/(N-M-P-1)$, we obtain the criteria

$$\text{AICc}(Y) = N\log|\hat{\boldsymbol{\Sigma}}_{YY}| + NP\log(2\pi) + N(N+1)\left(\frac{P}{N-P-2}\right), \tag{46}$$

$$\text{AICc}(X) = N\log|\hat{\boldsymbol{\Sigma}}_{XX}| + NM_K\log(2\pi) + N(N+1)\left(\frac{M_K}{N-M_K-2}\right), \tag{47}$$

$$\text{AICc}(XY) = N\log|\hat{\boldsymbol{\Sigma}}_{(XY)}| + N(P+M_K)\log(2\pi) + N(N+1)\left(\frac{M_K+P}{N-M_K-P-2}\right). \tag{48}$$

Conditional independence can be expressed in many different ways. A criterion for conditional independence should make consistent decisions for equivalent formulations. While such consistency is guaranteed for population quantities like cross entropy, it is not guaranteed for sample criteria. The following proposition gives the necessary condition for a sample criterion to give consistent decisions about conditional independence.

Proposition 6. Let $\mathcal{AIC}(Y|X) = N\log|\hat{\boldsymbol{\Sigma}}_{Y|X}| + \mathcal{P}$ be a sample criterion (note that AICc is of this form). Define the associated chain rule to be $\mathcal{AIC}(XY) = \mathcal{AIC}(X) + \mathcal{AIC}(Y|X)$. If $\mathcal{AIC}(Y|X)$ satisfies the chain rule, then it makes consistent decisions about $Y \perp X_R \,|\, X_K$. If it violates the chain rule, then there exists a sample for which it makes contradictory decisions about $Y \perp X_R \,|\, X_K$.

Proof. Let ω denote the constraint $Y \perp X_R \,|\, X_K$. Therefore, the associated candidate PDF $q_\omega(\cdot)$ satisfies (24)–(27). Based on these identities, the positivity of the following quantities are equally valid criteria for deciding ω:

$$\hat{\delta}_1 = \mathcal{AIC}(Y|X_KX_R) - \mathcal{AIC}_\omega(Y|X_KX_R) = \mathcal{AIC}(Y|X_KX_R) - \mathcal{AIC}(Y|X_K), \tag{49}$$

$$\hat{\delta}_2 = \mathcal{AIC}(X_R|X_KY) - \mathcal{AIC}_\omega(X_R|X_KY) = \mathcal{AIC}(X_R|X_KY) - \mathcal{AIC}(X_R|X_K), \tag{50}$$

$$\hat{\delta}_3 = \mathcal{AIC}(YX_R|X_K) - \mathcal{AIC}_\omega(YX_R|X_K) = \mathcal{AIC}(YX_R|X_K) - (\mathcal{AIC}(Y|X_K) + \mathcal{AIC}(X_R|X_K)), \tag{51}$$

$$\hat{\delta}_4 = \mathcal{AIC}(YX_KX_R) - \mathcal{AIC}_\omega(YX_KX_R) = \mathcal{AIC}(YX_KX_R) - (\mathcal{AIC}(YX_K) + \mathcal{AIC}(X_KX_R) - \mathcal{AIC}(X_K)). \tag{52}$$

If $\mathcal{AIC}$ satisfies the chain rule, then a little algebra shows $\hat{\delta}_1 = \hat{\delta}_2 = \hat{\delta}_3 = \hat{\delta}_4$; hence, $\mathcal{AIC}$ gives consistent decisions about $Y \perp X_R \,|\, X_K$. Note that $\hat{\delta}_i$ is of the form

$$\hat{\delta}_i = N\log\Lambda_K + \delta\mathcal{P}_i,$$

where $\delta\mathcal{P}_i$ is a positive, deterministic term that depends only on (N, M_K, M_R, P) and Λ_K is independent of i because by (45)

$$\Lambda_K = \frac{|\hat{\boldsymbol{\Sigma}}_{Y|X_KX_R}|}{|\hat{\boldsymbol{\Sigma}}_{Y|X_K}|} = \frac{|\hat{\boldsymbol{\Sigma}}_{X_R|X_KY}|}{|\hat{\boldsymbol{\Sigma}}_{X_R|X_K}|} = \frac{|\hat{\boldsymbol{\Sigma}}_{(YX_R)|X_K}|}{|\hat{\boldsymbol{\Sigma}}_{Y|X_K}||\hat{\boldsymbol{\Sigma}}_{X_R|X_K}|} = \frac{|\hat{\boldsymbol{\Sigma}}_{(YX)}||\hat{\boldsymbol{\Sigma}}_{X_KX_K}|}{|\hat{\boldsymbol{\Sigma}}_{(YX_K)}||\hat{\boldsymbol{\Sigma}}_{XX}|}. \tag{53}$$

In fact, Λ_K is a likelihood ratio because $\hat{\delta}_1, \hat{\delta}_2, \hat{\delta}_3, \hat{\delta}_4$ are nested comparisons. Therefore, Λ_K is a random variable on (0, 1]. Suppose $\mathcal{AIC}$ violates the chain rule; hence, for some sample, $\hat{\delta}_i \neq \hat{\delta}_j$. Then because Λ_K does not depend on i, $\delta\mathcal{P}_i \neq \delta\mathcal{P}_j$ for the parameters (N, M_K, M_R, P) of that sample. Because $-\log\Lambda_K$ is a continuous random variable with positive support on $[0, \infty)$, there is nonzero probability that it lies between $\delta\mathcal{P}_i$ and $\delta\mathcal{P}_j$. When this occurs, $\hat{\delta}_i$ and $\hat{\delta}_j$ have opposite signs and therefore $\mathcal{AIC}$ gives contradictory decisions about $Y \perp X_R \,|\, X_K$.

Unfortunately, AICc does *not* satisfy the chain rule; that is, $\text{AICc}(XY) \neq \text{AICc}(X) + \text{AICc}(Y|X)$. The reason AICc violates the chain rule is because its derivation implicitly assumes $\mathbf{X}_0 = \hat{\mathbf{X}}$ (DelSole & Tippett, 2021; Tian et al., 2020), which contradicts the assumption in (38) that $(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ and $(\mathbf{X}_0, \mathbf{Y}_0)$ are independent. Following Rosset and Tibshirani (2020), we define the following.

Definition 5. $\mathbf{X}_0$ and $\hat{\mathbf{X}}$ are said to be Same-X if $\mathbf{X}_0 = \hat{\mathbf{X}}$.

Definition 6. $\mathbf{X}_0$ and $\hat{\mathbf{X}}$ are said to be Random-X if the rows of $\mathbf{X}_0$ and $\hat{\mathbf{X}}$ are independently and identically distributed as a joint normal distribution.

AICc is an unbiased estimate of $\mathbb{IC}$ for Same-X. An important special case of Same-X is the intercept-only models (43). In this case, $\text{AICc}(Y), \text{AICc}(X), \text{AICc}(XY)$ still are the correct unbiased estimates of $\mathbb{IC}(Y), \mathbb{IC}(X), \mathbb{IC}(XY)$, because the only explanatory variable in each model is the intercept, which is Same-X, and therefore consistent with the derivation of Hurvich and Tsai (1989). However, under Random-X, $\text{AICc}(Y|X)$ violates the chain rule, and therefore, by Proposition 6, AICc can make contradictory decisions about $Y \perp X_R \,|\, X_K$. For these reasons, AICc is unsuitable for selecting models under Random-X. The appropriate sample criterion for Random-X is given in the next proposition.

Proposition 7. Assuming the candidate model (41) includes the true model, an unbiased estimate of $\mathbb{IC}(Y|X)$ under Random-X is

$$\text{AICr}(Y|X) = N \log|\hat{\boldsymbol{\Sigma}}_{Y|X}| + NP\log(2\pi) + N(N+1)\left(\frac{M_K+P}{N-M_K-P-2} - \frac{M_K}{N-M_K-2}\right). \tag{54}$$

Proof. Under Random-X, $\mathbb{IC}$ satisfies the chain rule (39); therefore, $\mathbb{IC}(Y|X)$ can be estimated as $\mathbb{IC}(XY) - \mathbb{IC}(X)$. Unbiased estimates of the latter two matrices are (48) and (47), respectively. Taking the difference $\text{AICc}(XY) - \text{AICc}(X)$ yields (54). Alternatively, $\text{AICr}(Y|X)$ can be derived by exact integration, as shown in DelSole and Tippett (2021) (see also Fujikoshi, 1985; Tian et al., 2020). AICr is written in the form (54), rather than in other forms in DelSole and Tippett (2021), to facilitate comparisons discussed below.

AICr satisfies the chain rule $\text{AICr}(XY) = \text{AICr}(X) + \text{AICr}(Y|X)$, and hence by Proposition 6, it gives consistent decisions for equivalent selection problems. Since AICr also is an unbiased estimate of $\mathbb{IC}$ for Random-X, it is the natural basis for estimating Akaike's extension of $\mathbb{MIC}$.

Proposition 8. Assuming the candidate PDF (41) includes the true PDF, an unbiased estimate of $\mathbb{MIC}a(X;Y)$ under Random-X is

$$\text{MIC}(X;Y) = \text{AICr}(Y|X) - \text{AICr}(Y) \tag{55}$$

$$= \text{AICr}(X|Y) - \text{AICr}(X) \tag{56}$$

$$= \text{AICr}(XY) - \text{AICr}(X) - \text{AICr}(Y). \tag{57}$$

In terms of the regression model (41),

$$\text{MIC}(X;Y) = N\log|\hat{\boldsymbol{\Sigma}}_{Y|X}| - N\log|\hat{\boldsymbol{\Sigma}}_{YY}| + \mathcal{P}(N,M_K,P), \tag{58}$$

where

$$\mathcal{P}(N,M_K,P) = N(N+1)\left(\frac{M_K+P}{N-M_K-P-2} - \frac{M_K}{N-M_K-2} - \frac{P}{N-P-2}\right). \tag{59}$$

Proof. Equation (55) follows from Proposition 5 after replacing $\mathbb{IC}$ with the estimate AICr. Equations (56) and (57) follow from (55) because AICr satisfies the chain rule. Equation (58) follows from (55) and (54).

Proposition 9 Sample criterion for *X*-selection. Under $\omega : Y \perp X_R \mid X_K$, (27) implies that

$$\text{AICr}_\omega(YX_RX_K) = \text{AICr}(YX_K) + \text{AICr}(X_RX_K) - \text{AICr}(X_K), \tag{60}$$

and therefore, a criterion for ω is $\Delta_X < 0$, where

$$\Delta_X = \text{MIC}(X_KX_R;Y) - \text{MIC}(X_K;Y) = \text{AICr}(YX_KX_R) - \text{AICr}_\omega(YX_KX_R). \tag{61}$$

Proposition 10 Sample criterion for simultaneous selection. Under $\psi : Y_K \perp X_R | X_K \text{and} Y_R \perp X_K X_R | Y_K$, (35) implies that

$$\text{AICr}_\psi(X_K, X_R, Y_K, Y_R) = \text{AICr}(Y_R|Y_K) + \text{AICr}(Y_K|X_K) + \text{AICr}(X_K, X_R), \tag{62}$$

and therefore, a criterion for ψ is $\Delta_{XY} < 0$, where

$$\Delta_{XY} = \text{MIC}(X_K X_R; Y_K Y_R) - \text{MIC}(X_K; Y_K) = \text{AICr}(Y_R X_R Y_K X_K) - \text{AICr}_\psi(Y_R X_R Y_K X_K). \tag{63}$$

Proposition 11. Partition the matrices in (41) as $\mathbf{X} = [\mathbf{X}_K \mathbf{X}_R]$ and $\mathbf{Y} = [\mathbf{Y}_K \mathbf{Y}_R]$, where $\mathbf{X}_K, \mathbf{X}_R, \mathbf{Y}_K, \mathbf{Y}_R$ are each full column rank matrices of rank M_K, M_R, P_K, P_R, respectively, with $M = M_K + M_R$ and $P = P_K + P_M$. Then $\Delta_X = N \log\Lambda_K + \mathcal{P}(N, M_K + M_R, P) - \mathcal{P}(N, M_K, P)$, or

$$\Delta_X = N \log\Lambda_K + N(N+1)\left(\frac{M_K + M_R + P}{N - M_K - M_R - P - 2} - \frac{M_K + M_R}{N - M_K - M_R - 2} - \frac{M_K + P}{N - M_K - P - 2} + \frac{M_K}{N - M_K - 2}\right).$$

Similarly, $\Delta_{XY} = N \log|\hat{\boldsymbol{\Sigma}}_{(Y_R X_R)|(Y_K X_K)}| - N \log|\hat{\boldsymbol{\Sigma}}_{X_R|X_K}| - N \log|\hat{\boldsymbol{\Sigma}}_{Y_R|Y_K}| + \mathcal{P}(N, M_K + M_R, P_K + P_R) - \mathcal{P}(N, M_K, P_K)$, or equivalently

$$\begin{aligned}\Delta_{XY} = {} & N \log|\hat{\boldsymbol{\Sigma}}_{(Y_R X_R)|(Y_K X_K)}| - N \log|\hat{\boldsymbol{\Sigma}}_{X_R|X_K}| - N \log|\hat{\boldsymbol{\Sigma}}_{Y_R|Y_K}| \\ & + N(N+1)\left[\frac{M+P}{N-P-M-2} - \frac{M}{N-M-2} - \frac{P}{N-P-2} - \frac{M_K + P_K}{N - M_K - P_K - 2} + \frac{M_K}{N - M_K - 2} + \frac{P_K}{N - P_K - 2}\right].\end{aligned} \tag{64}$$

Remark 1. Many standard texts recommend using AICc for X-selection (e.g., Burnham & Anderson, 2002). We argue that AICc is not suitable for deciding conditional independence because it gives inconsistent decisions for equivalent formulations of conditional independence. Another issue can be seen by comparing Δ_X to $\Delta'_X = \text{AICc}(Y|X_K X_R) - \text{AICc}(Y|X_K)$. The latter criterion imposes less penalty per each extra predictor than does (61). The reason for this is that AICc assumes Same-X while AICr assumes Random-X (as discussed earlier in this section). As a result, AICc neglects a source of uncertainty and therefore underestimates the cross entropy.

Remark 2. Under normality, deciding $Y \perp X_R \mid X_K$ is equivalent to deciding

$$\mathbf{B}_R = \mathbf{0} \;\text{ in }\; \mathbf{Y} = \mathbf{X}_K \mathbf{B}_K + \mathbf{X}_R \mathbf{B}_R + \mathbf{j}\boldsymbol{\mu}_Y^T + \mathbf{E}_Y. \tag{65}$$

The likelihood ratio test (LRT Johnson & Wichern, 2002) for this hypothesis is to decide $\mathbf{B}_R = \mathbf{0}$ when $\Delta_{\text{LRT}} = \log\Lambda_K - \log\Lambda_C > 0$, where Λ_K is defined in (53), and Λ_C is the critical value from Wilks' lambda distribution with parameters $(P, M_R, N - M)$. Both Δ_{LRT} to Δ_X depend on sample values only through the likelihood ratio and therefore differ only by the critical value. However, the LRT is limited to nested models.

Remark 3. Conditional independence $\omega : Y \perp X_R \mid X_K$ also can be expressed as (26), which under normal distributions is equivalent to

$$\boldsymbol{\Sigma}^{\omega}_{YX_R|X_K} = \mathbf{0} \;\Leftrightarrow\; \boldsymbol{\Sigma}^{\omega}_{YX_R} = \boldsymbol{\Sigma}_{YX_K} \boldsymbol{\Sigma}^{-1}_{X_K X_K} \boldsymbol{\Sigma}_{X_K X_R}.$$

This is precisely the covariance constraint used by Fujikoshi et al. (2010) to derive a criterion for selecting one set of variables in CCA (see their sec. 11.5). The fact that this selection problem is equivalent to deciding ω indicates that a separate derivation is unnecessary.

Remark 4. Conditional independence $\omega : Y \perp X_R \mid X_K$ also can be expressed as (25), which under normal distributions is equivalent to the hypothesis $\mathbf{B}_Y = \mathbf{0}$ in the model

$$\mathbf{B}_Y = \mathbf{0} \text{ in } \mathbf{X}_R = \mathbf{Y}\mathbf{B}_Y + \mathbf{X}_K \mathbf{B}_K + \mathbf{E}.$$

Because this hypothesis is equivalent to ω, the criterion also is the same, as also can be seen from the following identity:

$$\mathrm{MIC}(X_R;YX_K)-\mathrm{MIC}(X_R;X_K)=\mathrm{AICr}(X_R|YX_K)-\mathrm{AICr}(X_R|X_K)=\mathrm{AICr}(Y|X_KX_R)-\mathrm{AICr}(Y|X_K).$$

Remark 5. Turning to an apparently different selection problem, Fujikoshi (1989) proposed a criterion for selecting Y variables on the basis that $\mathbf{y}_R$, after removing the effects of $\mathbf{y}_K$, does not depend on $\mathbf{x}$. This criterion can be framed as the hypothesis

$$\mathbf{B}_X=\mathbf{0}\ \text{ in }\ \mathbf{Y}_R=\mathbf{Y}_K\mathbf{B}_K+\mathbf{X}_K\mathbf{B}_X+\mathbf{j}\boldsymbol{\mu}_R^T+\mathbf{E}_R. \tag{66}$$

Under normality, the selection problem (66) is equivalent to deciding

$$X_K \perp \mathrm{Y_R}|\mathrm{Y_K}, \tag{67}$$

which is merely (23), except with X and Y labels switched. We call this *Y-selection*. Thus, all of the above results for X-selection can be applied immediately to Y -selection, after swapping variable labels. In particular, the criterion for Y -selection is

$$\mathrm{MIC}(X_K;Y_KY_R)-\mathrm{MIC}(X_K;Y_K)=N\log\left(\frac{|\hat{\boldsymbol{\Sigma}}_{Y_KY_R|X_K}|}{|\hat{\boldsymbol{\Sigma}}_{Y_KY_R}|}\right)-N\log\left(\frac{|\hat{\boldsymbol{\Sigma}}_{Y_K|X_K}|}{|\hat{\boldsymbol{\Sigma}}_{Y_K}|}\right)+\mathcal{P}(N,M_K,P_K+P_R)-\mathcal{P}(N,M_K,P_K). \tag{68}$$

This small-sample criterion is asymptotically equivalent to the criterion derived by Fujikoshi (1989). Because MIC is symmetric, the criterion is identical to regression model selection but with the usual roles of X and Y swapped; namely, X is response and Y is explanatory. In this sense, selecting response variables is fundamentally equivalent to selecting explanatory variables—once a criterion for X-selection exists, one can swap X and Y labels and apply it to select response variables. In this sense, a separate derivation of a criterion for Y -selection is unnecessary.

5 | MAXIMIZING THE LIKELIHOOD UNDER CONDITIONAL INDEPENDENCE CONSTRAINTS

It should be recognized that the criteria stated in Propositions 9 and 10 were obtained merely by evaluating MIC. In particular, no constrained maximum likelihood problem needed to be solved. Nevertheless, (61) and (63) assert that the criteria are equivalent to the AICr of the joint PDFs constrained by the relevant form of conditional independence. These assertions can be verified because the associated constrained optimization problems have in fact been solved in the literature, though this fact seems not to be widely recognized. First, Fujikoshi (1985) derived the corrected AIC criterion for X-selection under Random-X. The result is his eq. 5.17, which is identical to our $-\Delta_X$. This serves as a check on our derivation of (61). Also, this equivalence implies that Fujikoshi (1985) derived the small-sample correction to AIC under Random-X nearly 40 years ago!

In regards to Proposition 10, the verification is somewhat more complicated because the small-sample corrected AIC for simultaneous selection does not appear in the literature. However, Fujikoshi et al. (2010) derived a criterion based on Distance Information Criterion, which is closely related to AIC (see sec. 10.6.1 of Fujikoshi et al., 2010). The small-sample corrected version of this criterion is called CDIC and appears in sec. 11.5.2 of Fujikoshi et al. (2010). To remove the slight inconsistency with AIC, we adjust CDIC as follows: replace the overall factor of "n" by "N," and replace "n" in the numerator of each penalty term by "$N+1$," which yields the following modified criterion CDIC*:

$$\begin{aligned}\mathrm{CDIC}^* &= -N\log|\hat{\boldsymbol{\Sigma}}_{(Y_RX_R)|(Y_KX_K)}|+N\log|\hat{\boldsymbol{\Sigma}}_{X_R|X_K}|+N\log|\hat{\boldsymbol{\Sigma}}_{Y_R|Y_K}|\\ &\quad+N\left(\frac{(N+1)(M_K+P_K)}{N-M_K-P_K-2}+\frac{(N+1)(M_K+M_R)}{N-M_K-M_R-2}+\frac{(N+1)(P_K+P_R)}{N-P_K-P_R-2}-\frac{(N+1)(M_K)}{N-M_K-2}-\frac{(N+1)(P_K)}{N-P_K-2}-(P+M-1)\right).\end{aligned}$$

Comparison with (64) shows that CDIC^* and $-\Delta_{XY}$ agree, except for additive terms that depend only on N and $P+M$. It is not clear why there exist differing terms, but Fujikoshi et al. (2010) applied their criterion to situations in which N and $P+M$ were constant; hence, these terms do not affect model selection. We interpret this agreement as confirming that both Δ_X and Δ_{XY} are the correct small-sample criteria for conditional independence.

Importantly, Fujikoshi (1985) and Fujikoshi et al. (2010) derived the above criteria by explicitly maximizing the likelihood function subject to a constraint associated with conditional independence. The solution to such constrained optimization problems requires intricate matrix manipulations. In contrast, the criteria in Proposition 11 were obtained simply by taking differences in MIC. The simplicity in the latter approach derives

from the fact that certain forms of conditional independence allow structured PDFs to be expressed in terms of criteria for unstructured PDFs. Specifically, the left-hand sides of (60) and (62) require solving a constrained ML problem whereas the right-hand side requires solving unconstrained ML problems. A remarkable fact is that MIC gives this decomposition directly simply by computing differences in MIC of appropriate variable subsets.

6 | CANONICAL CORRELATION ANALYSIS

MIC is a natural criterion for CCA because, in addition to the above reasons, it depends on sample values *only* through the canonical correlations.

Proposition 12. Let the canonical correlations between **X** and **Y** in (41) be $\hat{\rho}_1, \hat{\rho}_2, \ldots$. Then

$$\mathrm{MIC}(Y;X) = N\sum_i \log(1-\hat{\rho}_i^2) + \mathcal{P}(N, M_K, P). \tag{69}$$

Proof. Recall that canonical correlations are derived from the eigenvalues of

$$\hat{\boldsymbol{\Sigma}}_{YX}\hat{\boldsymbol{\Sigma}}_{XX}^{-1}\hat{\boldsymbol{\Sigma}}_{XY}\mathbf{w}_Y = \hat{\rho}^2\hat{\boldsymbol{\Sigma}}_{YY}\mathbf{w}_Y \;\Leftrightarrow\; \hat{\boldsymbol{\Sigma}}_{Y|X}\mathbf{w}_Y = (1-\hat{\rho}^2)\hat{\boldsymbol{\Sigma}}_{YY}\mathbf{w}_Y,$$

where (44) has been used. Since the determinant of a matrix equals the product of eigenvalues,

$$\left|\hat{\boldsymbol{\Sigma}}_{YY}^{-1}\hat{\boldsymbol{\Sigma}}_{Y|X}\right| = \prod_i (1-\hat{\rho}_i^2).$$

Taking the log of both sides and substituting the result into (58) yields (69).

For normal distributions, a sample estimate of mutual information is (Soofi et al., 2010),

$$\mathbb{M}(X;Y)_{\text{Gaussian}} \approx -\frac{1}{2}\sum_i \log(1-\hat{\rho}_i^2). \tag{70}$$

Thus, minimizing MIC strikes a balance between maximizing mutual information while minimizing the number of parameters being estimated. To illustrate the application of MIC for selecting variables in CCA, consider data generated by the model

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \epsilon, \tag{71}$$

where $M_K = 10$, $P = 10$, $\mathbf{A} = \rho\mathbf{v}\mathbf{v}^T$, $\mathbf{v} = (1\ 2\ 2\ 1\ 0\ 0\ 0\ 0\ 0\ 0)^T/\sqrt{10}$, $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$, $\mathbf{Q} = \mathbf{I} - \rho\mathbf{v}\mathbf{v}^T$, $\rho = 0.7$. When all 10 X and Y variables are included, population CCA yields one nonzero canonical correlation, namely, $\rho_1 = 0.7$. However, only the first four X and first four Y variables are relevant; additional variables beyond this add no information about the X–Y relation and are therefore redundant. We consider a selection problem in which the candidate variables are included in a sequentially nested fashion; that is, the candidate model with M_K X variables consists of $X_1, X_2, \ldots, X_{M_K}$, and the candidate model with M_Y Y variables consists of $Y_1, Y_2, \ldots, Y_{M_Y}$.

Figure 1 shows MIC for a particular realization of samples for $N = 50$. The minimum MIC occurs when three X and three Y variables are used. Repeating this procedure 100 times and counting the number of times a particular model is selected leads to the top left panel of Figure 2. For reference, the population mutual information is indicated by the shading. The most common selection is for three X and three Y variables. For comparison, we define an "uncorrected MIC" using (58) but with the uncorrected penalty $\lim_{N\to\infty}\mathcal{P}(N, M_K, P) = 2M_K P$. Selections based on uncorrected MIC, shown in the bottom left panel, show much larger tendency to overfit, which illustrates the importance of using the corrected criterion. For a larger sample size, $N = 200$ (right column), MIC overwhelmingly selects four X and four Y variables, the correct choice for large N. The uncorrected MIC still shows a larger tendency to overfit. Even for $N = 20{,}000$ (not shown), MIC overwhelming selects four X and four Y variables and shows little tendency to overfit.

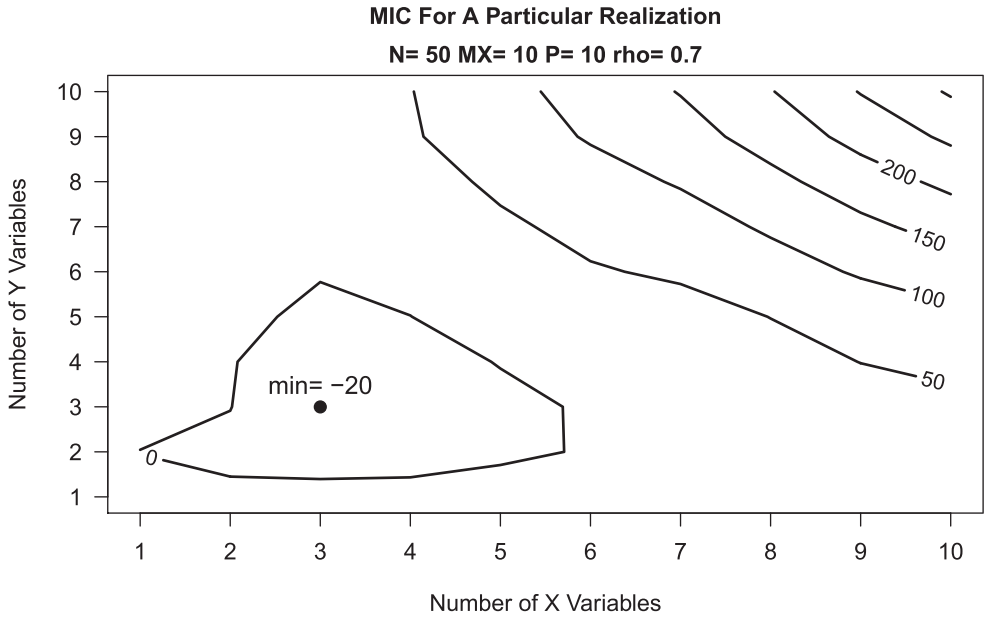

FIGURE 1 Contours of MIC for a particular realization from the model (71). The minimum MIC is indicated by a dot and is labelled

7 | GRAPHICAL MODELS

We now consider using MIC to select graphical models. Graphical models express conditional dependencies by a graph comprising nodes and edges, where the absence of an edge between two nodes indicates that those two variables are conditionally independent given all other variables. More precisely, the two nodes Z_1 and Z_2 have no edge if

$$\omega_{12} : Z_1 \perp Z_2 | Z_{/12}, \tag{72}$$

where $Z_{/12}$ means "all Z-variables except Z_1 and Z_2." Graphical models corresponding to *X*-selection, *Y*-selection, and simultaneous selection are illustrated in Figure 3. The graph for simultaneous selection follows from the fact that

$$Y_R \perp X_K X_R | Y_K \Rightarrow \begin{cases} Y_R \perp X_K | (Y_K X_R) \\ Y_R \perp X_R | (Y_K X_K) \end{cases}, \tag{73}$$

(which follows from the converse of Lemma 4.3 in Dawid, 1979). The associated structures have a simple expression in terms of the precision matrix (i.e., the inverse of the covariance matrix). Specifically, (72) implies that the (Z_1, Z_2) element of the precision matrix vanishes. Accordingly, the precision matrices corresponding to *X*-selection, *Y*-selection, and simultaneous selection have, respectively, the following forms

$$\begin{array}{c} \\ Y_K \\ X_K \\ X_R \end{array} \begin{array}{c} \begin{array}{ccc} Y_K & X_K & X_R \end{array} \\ \begin{pmatrix} \times & \times & 0 \\ \times & \times & \times \\ 0 & \times & \times \end{pmatrix} \end{array}, \quad \begin{array}{c} \\ Y_R \\ Y_K \\ X_K \end{array} \begin{array}{c} \begin{array}{ccc} Y_R & Y_K & X_K \end{array} \\ \begin{pmatrix} \times & \times & 0 \\ \times & \times & \times \\ 0 & \times & \times \end{pmatrix} \end{array}, \quad \begin{array}{c} \\ Y_R \\ Y_K \\ X_K \\ X_R \end{array} \begin{array}{c} \begin{array}{cccc} Y_R & Y_K & X_K & X_R \end{array} \\ \begin{pmatrix} \times & \times & 0 & 0 \\ \times & \times & \times & 0 \\ 0 & \times & \times & \times \\ 0 & 0 & \times & \times \end{pmatrix} \end{array},$$

A standard result in information theory is that if ω_{12} is true, then conditional mutual information vanishes; that is, $\mathbb{M}(Z_1;Z_2|Z_{/12}) = 0$. As remarked in (20), a conditional $\mathbb{MIC}$ may be defined that behaves analogously to conditional mutual information, except it varies in the opposite way (i.e., large $\mathbb{MIC}$ corresponds to weak conditional independence). By suitable redefinition of variable labels in previous sections, conditional $\mathbb{MIC}$ is

$$\mathbb{MIC}(Z_1;Z_2|Z_{/12}) = \mathbb{MIC}(Z_1;Z_{/1}) - \mathbb{MIC}(Z_1;Z_{/12}) = \mathbb{H}(Z_1|Z_{/1}) - \mathbb{H}(Z_1|Z_{/12}) = \mathbb{H}(Z) - \mathbb{H}_{\omega_{12}}(Z). \tag{74}$$

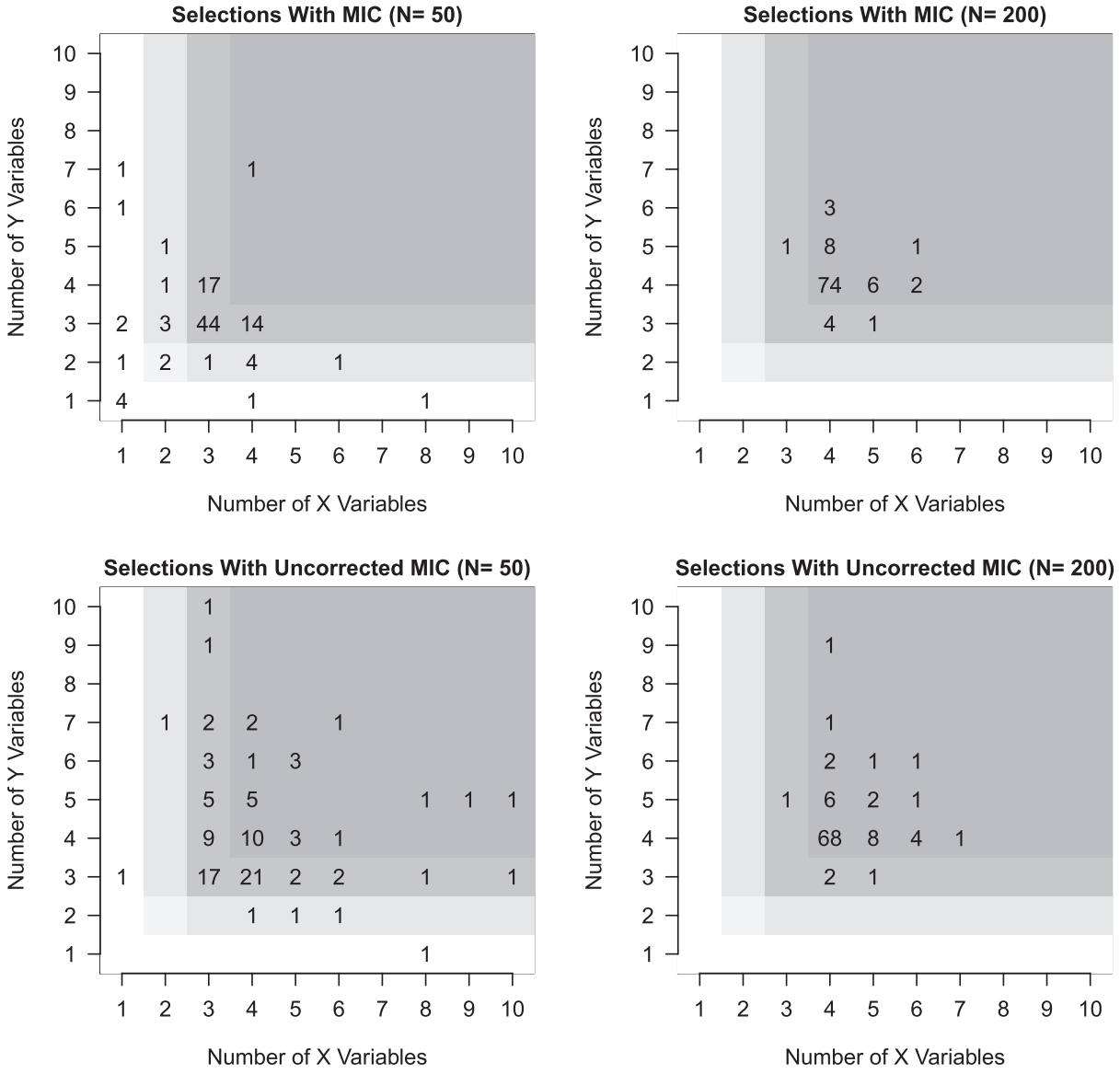

FIGURE 2 Number of times MIC selects the number of X and Y variables for CCA for 100 independent realizations from the model (71). Results are shown for samples sizes $N = 50$ (left column) and $N = 200$ (right column), and using MIC (top row) and uncorrected MIC (bottom row). The shading shows the population mutual information (it is the same in all panels)

The Akaike-based sample estimate of conditional MIC is

$$\mathrm{MIC}(Z_1;Z_2|Z_{/12}) = \mathrm{MIC}(Z_{/1};Z_1) - \mathrm{MIC}(Z_{/12};Z_1).$$

The decision rule is to accept ω_{12} if $\mathrm{MIC}(Z_1;Z_2|Z_{/12}) > 0$. This criterion can be evaluated for any Z_1 and Z_2, even if the graph is nondecomposable. For completeness, we note that the analogous criterion for deciding $Z_1 \perp Z_2$ is $\mathbb{MIC}(Z_1;Z_2) = \mathbb{H}(Z_1|Z_2) - \mathbb{H}(Z_1) > 0$.

Proposition 13. For scalar Z_1 and Z_2, conditional MIC is

$$\begin{aligned}\mathrm{MIC}(Z_1;Z_2|Z_{/12}) &= \log\left(\frac{|\boldsymbol{\Sigma}_{Z_1Z_2|Z_{/12}}|}{|\boldsymbol{\Sigma}_{Z_1|Z_{/12}}||\boldsymbol{\Sigma}_{Z_2|Z_{/12}}|}\right) + \mathcal{P}(N,1,D-1) - \mathcal{P}(N,1,D-2) \\ &= \log\left(1-\hat{\rho}^2_{12|Z_{/12}}\right) + \mathcal{P}(N,1,D-1) - \mathcal{P}(N,1,D-2),\end{aligned} \tag{75}$$

where D is the total number of Z-variables and $\hat{\rho}_{12|Z_{/12}}$ is the partial correlation between Z_1 and Z_2 after regressing out $Z_{/12}$.

One of the most popular algorithms for identifying graphical models is the PC Algorithm (Spirtes et al., 2001). This algorithm requires a criterion for deciding conditional independence. A standard criterion is based on statistical significance of the partial correlation. However, significance depends on the arbitrary significance level α and is not guaranteed to be a proper score. In contrast, the criterion $\mathrm{MIC}(Z_1;Z_2|Z_{/12}) > 0$ does not depend on the arbitrary α and is a proper score. To illustrate its application, we consider a simple four-variable model governed by

$$Y_R = AY_K + E_1,\ Y_K = BX_K + E_2,\ X_R = CX_K + E_3,\ X_K = E_4, \tag{76}$$

which corresponds to the right-most graph in Figure 3 for simultaneous selection. Our goal here is not to derive a new algorithm for exploring the space of all graphs but rather to illustrate the impact of using a different criterion for deciding conditional independence. Accordingly, we have used the `pcalg` package in R to select the graph from samples generated by model (76). The values of A, B, C were generated randomly from $\mathcal{N}(2,1)$. Then, for the selected A, B, C, we generated N samples from (76), where E_1, E_2, E_3, E_4 are independently drawn from $\mathcal{N}(0,1)$. Then, this whole procedure (including resampling A, B, C) was repeated 1000 times, and the number of times the PC algorithm identified the correct graph was recorded. We have performed two different experiments: one that decides conditional independence based on significance of the partial correlation using $\alpha = 5\%$, and one based on MIC in (75). The results are shown in Figure 4. The figure shows that for this choice of α and for a small

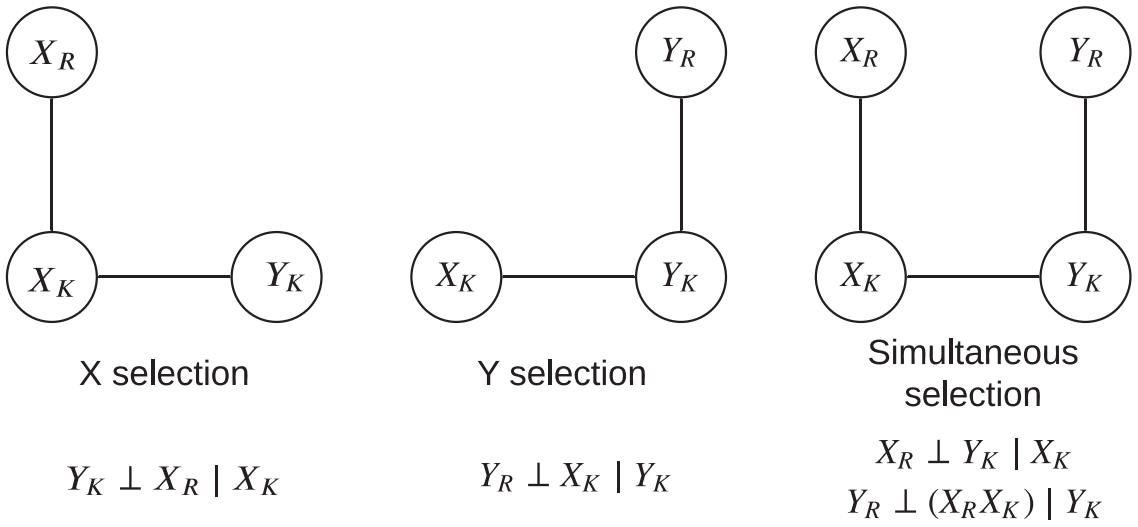

FIGURE 3 Graphical models associated with *X*-selection, *Y*-selection, and simultaneous selection

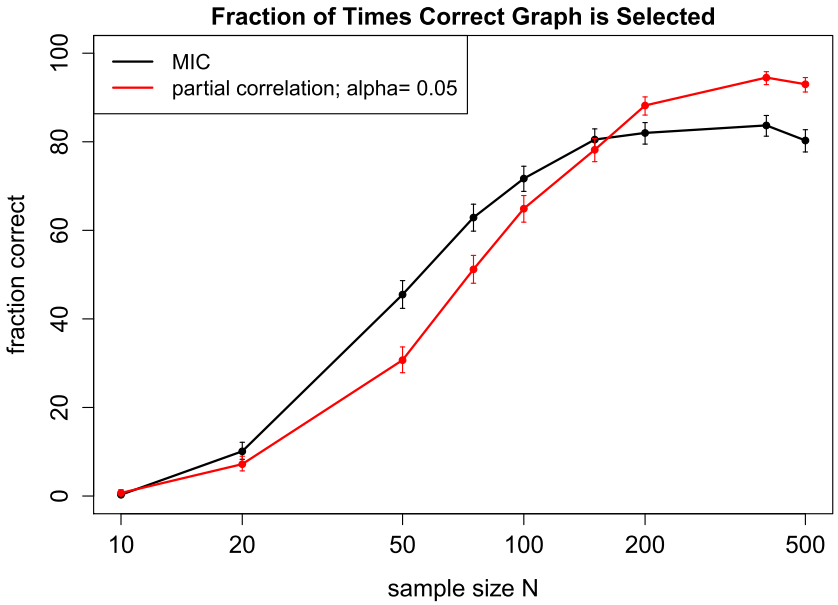

FIGURE 4 Fraction of times the PC Algorithm selects the correct graph from model (76), whose graph is the right-most graph in Figure 3 corresponding to simultaneous selection, as a function of sample size N. The PC algorithm is run in two modes, one using $M(Z_1; Z_2 \mid Z_{/12}) > 0$ from (75) (black), and one using significance of the partial correlation at the 5% level (red). The error bars show 95% confidence intervals based on 1000 trials

sample size (N < 100), the PC algorithm selects the correct graph more frequently using MIC than using the significance of the partial correlation. This result should not be interpreted as general, since it depends on the choice of α, which is a tuning parameter in the PC algorithm. In contrast, the criterion $\mathrm{MIC}(Z_1;Z_2|Z_{/12}) > 0$ does not involve tunable parameters. The parameter α could be tuned to produce better results, but this tuning is not generally possible when the true graph is unknown. We emphasize that the criterion for conditional independence (74) is not restricted for univariate Z_1 and Z_2; hence, this criterion may open new approaches to graphical model selection.

ACKNOWLEDGEMENTS

This research was supported primarily by the National Science Foundation (AGS-1822221). Additional support was provided from National Science Foundation (AGS-1338427), National Aeronautics and Space Administration (NNX14AM19G), and the National Oceanic and Atmospheric Administration (NA14OAR4310160). The views expressed herein do not necessarily reflect the views of these agencies.

DATA AVAILABILITY STATEMENT

No original data were generated through this work.

ORCID

Timothy DelSole https://orcid.org/0000-0003-2041-3024
Michael K. Tippett https://orcid.org/0000-0002-7790-5364

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N., & Czáki, F. (Eds.), *2nd International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiadó.

Barndorff-Nielsen, O. (1976). Factorization of likelihood functions for full exponential families. *Journal of the Royal Statistical Society. Series B (Methodological)*, *38*(1), 37–44. http://www.jstor.org/stable/2984826

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, *41*(1), 1–31.

DelSole, T., & Tippett, M. K. (2021). Correcting the corrected AIC. *Statistics & Probability Letters*, *173*, 109064. https://www.sciencedirect.com/science/article/pii/S0167715221000262

Fan, J., Feng, Y., & Xia, L. (2020). A projection-based conditional dependence measure with applications to high-dimensional undirected graphical models. *Journal of Econometrics*, *218*(1), 119–139. https://www.sciencedirect.com/science/article/pii/S0304407620300403

Fujikoshi, Y. (1982). A test for additional information in canonical correlation analysis. *Annals of the Institute of Statistical Mathematics*, *34*(3), 523–530. https://doi.org/10.1007/BF02481050

Fujikoshi, Y. (1985). Selection of variables in discriminant analysis and canonical correlation analysis. In Krishnaiah, P. R. (Ed.), *Multivariate analysis VI: Proceedings of the Sixth International Symposium on Multivariate Analysis* (pp. 219–236). New York: Elsevier.

Fujikoshi, Y. (1989). Tests for redundancy of some variables in multivariate analysis. In Dodge, Y. (Ed.), *Statistical data analysis and inference* (pp. 141–163), Amsterdam: North-Holland. http://www.sciencedirect.com/science/article/pii/B9780444880291500186

Fujikoshi, Y., Ulyanov, V. V., & Shimizu, R. (2010). *Multivariate statistics: High-dimensional and large-sample approximations*. Hoboken, New Jersey: John Wiley and Sons.

Huang, T.-M. (2010). Testing conditional independence using maximal nonlinear conditional correlation. *The Annals of Statistics*, *38*(4), 2047–2091. https://doi.org/10.1214/09-AOS770

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297–307.

Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis* (5th ed.). Upper Saddle River, New Jersey: Prentice-Hall.

Rosset, S., & Tibshirani, R. J. (2020). From fixed-X to random-X regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association*, *115*(529), 138–151. https://doi.org/10.1080/01621459.2018.1424632

Soofi, E. S., Zhao, H., & Nazareth, D. L. (2010). Information measures. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*, 75–86.

Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, prediction, and search* (2nd ed.). Cambridge, Massachusetts: A Bradford Book.

Su, L., & White, H. (2007). A consistent characteristic function-based test for conditional independence. *Journal of Econometrics*, *141*(2), 807–834. https://www.sciencedirect.com/science/article/pii/S0304407606002375

Su, L., & White, H. (2008). A nonparametric Hellinger metric test for conditional independence. *Econometric Theory*, *24*(4), 829–864.

Sugiura, N. (1978). Further analysts of the data by Akaike' s information criterion and the finite corrections. *Communications in Statistics - Theory and Methods*, *7*(1), 13–26. https://doi.org/10.1080/03610927808827599

Tian, S., Hurvich, C. M., & Simonoff, J. S. (2020). Selection of regression models under linear restrictions for fixed and random designs. ArXiv e-prints, 28.

How to cite this article: DelSole, T., & Tippett, M. K. (2021). A mutual information criterion with applications to canonical correlation analysis and graphical models. *Stat*, *10*(1), e385. https://doi.org/10.1002/sta4.385