Estimating the Causal Effect from Partially Observed Time Series

Akane Iseki,[1] **Yusuke Mukuta,**[1,2] **Yoshitaka Ushiku,**[1] **Tatsuya Harada**[1,2]
[1]The University of Tokyo [2]RIKEN AIP
{iseki, mukuta, ushiku, harada}@mi.t.u-tokyo.ac.jp

Abstract

Many real-world systems involve interacting time series. The ability to detect causal dependencies between system components from observed time series of their outputs is essential for understanding system behavior. The quantification of causal influences between time series is based on the definition of some causality measure. Partial Canonical Correlation Analysis (Partial CCA) and its extensions are examples of methods used for robustly estimating the causal relationships between two multidimensional time series even when the time series are short. These methods assume that the input data are complete and have no missing values. However, real-world data often contain missing values. It is therefore crucial to estimate the causality measure robustly even when the input time series is incomplete. Treating this problem as a semi-supervised learning problem, we propose a novel semi-supervised extension of probabilistic Partial CCA called semi-Bayesian Partial CCA. Our method exploits the information in samples with missing values to prevent the overfitting of parameter estimation even when there are few complete samples. Experiments based on synthesized and real data demonstrate the ability of the proposed method to estimate causal relationships more correctly than existing methods when the data contain missing values, the dimensionality is large, and the number of samples is small.

Introduction

Understanding the interdependence of multiple time series is essential in a wide range of fields such as meteorology, finance, and robotics. The interactions between system components are dynamic and complicated. To analyze and predict system behavior, it is important to detect the dependent components from the output time series. We can detect the interdependence of multiple time series by quantifying their pair-wise interdependence. Causality measure, e.g., in terms of Granger Causality (Granger 1969) is a powerful approach for revealing such pair-wise interdependences. It quantifies, within a given pair of time series, the influence of the past value in one time series on the future value in the other. Among the many types of causality measures proposed in the literature, Partial CCA (Rao 1969) is a useful multivariate analysis method for estimating the causal effects between high-dimensional data robustly. While the original paper does not mention the relationship with causality, (Shibuya, Harada, and Kuniyoshi 2011) showed that by using the projections of two time series estimated by Partial CCA, we can calculate Transfer Entropy, which is a causality measure equivalent to Granger Causality when variables follow a Gaussian distribution (Shibuya, Harada, and Kuniyoshi 2009).

However, these methods require a large number of samples with no missing values. This scenario is not always realized with real data, where missing values often arise because of sensor deficiencies or privacy regulations. Also, the quantification of a causal effect between two time series requires triplets of variables, i.e., the past value in one time series and both the current and the past value in the other time series. So, it is difficult to obtain many samples in which all the variables are simultaneously observed. Moreover, when causal effects change dynamically, we need to estimate causal relationships over a short time span where the causality measure can be assumed to be constant. Additionally, real data tend to contain redundant information and become high-dimensional. Thus, it is necessary to estimate the causality measure robustly from high-dimensional small-sample time series in which some values are missing.

A naive way to deal with incomplete data is to ignore samples with missing values. However, complete samples cannot provide enough information to estimate parameters and lead to overfitting. Few methods for estimating causality address this problem, but several multivariate analysis methods have been proposed to compensate the lack of information. One of these adopts a Bayesian approach; another is a semi-supervised approach. Bayesian approaches do not use missing samples but provide a way to cope with small samples by adding prior information on parameters (Klami, Virtanen, and Kaski 2013; Bishop 1999). Partial CCA also has Bayesian extensions (Mukuta and Harada 2014). Semi-supervised approaches compensate for the lack of information by exploiting incomplete samples. (Ilin and Raiko 2010) combines both approaches for principal component analysis , but it does not provide enough basis for explaining why the results do not suffer from any bias caused by missing data.

In this work, we propose a causality estimation method that combines the semi-supervised and Bayesian approaches. We addressed the issue of missing data by using Rubin's missing model (Rubin 1974; 1975; 1976), which was originally proposed for randomized experiments and observational studies. We applied it to causality estimation from time series for the first time. In particular, we regard the missingness of samples as a probabilistic variable and incorporate it into the model of Bayesian Partial CCA. We prove that we can estimate unbiased parameters without explicitly modeling missing mechanism if samples are "missing completely at random" (MCAR) or "missing at random" (MAR). Still, our method can prevent the divergence of parameters even in the case of "missing not at random" (MNAR).

We experimentally demonstrate that our method can exploit samples with missing values to make the estimation stable even when the number of complete samples is small and existing methods result in overfitting. An experiment with simulated data was conducted using several types of data with various dimensionalities, sample sizes, and ratio of missing values. We also assessed whether performance depends on whether there is incompleteness in the cause variable, the outcome variable, or both variables. The experimental results confirm the good performance of the method, especially when the dimensionality of the data is smaller than the number of complete samples. We display the performance of our method, both in terms of preventing the divergence of parameters and of detecting spurious causality. Analysis shows how our method robustly estimates the parameter that plays an essential role in causality quantification. Applications to real data are demonstrated on a meteorological dataset and on a video dataset. Both examples show that our method can estimate causality among multiple time series by applying it to all pairs of time series.

The study proceeds as follows:

- We address the problem of estimating the causality measure from two multi-dimensional time series with missing values. To solve this problem, we propose a semi-supervised extension of Bayesian Partial CCA called semi-Bayesian Partial CCA (semi-BPCCA).
- Semi-BPCCA can estimate the causal effect independently of the probabilistic model of missing of data when the missingness is independent of the missing variable.
- Experiments on both artificial data and real data demonstrate that our method can avoid overfitting by exploiting non-paired samples.

Related Work

This section explains the existing method of causality quantification between two multi-dimensional time series.

Transfer Entropy and Partial CCA

Transfer entropy (TE) (Schreiber 2000) is the causality measure derived from the perspective of information theory. TE quantifies a causal effect as an information flow from one time series to the other. This measure is based on the assumption that the causal effect from time series $X = [x_1, ..., x_t, ..., x_T] \in \mathbb{R}^{d_x \times T}$ to series $Y = [y_1, ..., y_t, ..., y_T] \in \mathbb{R}^{d_y \times T}$ is large when past $x_{t-1}^{(k)}$ is predictive of the current y_t and can improve a Markov-model-based prediction of y_t using only the past value $y_{t-1}^{(l)}$. Here, $x_{t-1}^{(k)} = [x_{t-1}^T, ..., x_{t-k}^T]^T$ and $y_{t-1}^{(l)} = [y_{t-1}^T, ..., y_{t-l}^T]^T$. More specifically, $TE_{X\to Y}$ is defined as the Kullback Leibler divergence between conditional distributions of y_t only given $y_{t-1}^{(l)}$ and given $y_{t-1}^{(l)}$ and $x_{t-1}^{(k)}$ written as

$$\begin{aligned} TE_{X\to Y} &= \iiint p(y_t, y_{t-1}^{(l)}, x_{t-1}^{(k)}) \\ &\qquad \log_2 \frac{p(y_t|y_{t-1}^{(l)}, x_{t-1}^{(k)})}{p(y_t|y_{t-1}^{(l)})} dy_t dy_{t-1}^{(l)} dx_{t-1}^{(k)}. \quad (1)\end{aligned}$$

TE has a wide range of applications in, e.g., economics (Kwon and Yang 2008), robotics (Berger et al. 2016), and neuroscience (Wibral et al. 2013).

Partial CCA is a multivariate analysis method that calculates the projections that maximize the partial correlation between two input variables $y^1 \in \mathbb{R}^{d_{y^1}}$ and $y^2 \in \mathbb{R}^{d_{y^2}}$ after eliminating the effect of a third variable $x \in \mathbb{R}^{d_x}$. The projections a and b are calculated as the following generalized eigenvalue problem

$$\Sigma_{y^1y^2|x}\Sigma_{y^2y^2|x}^{-1}\Sigma_{y^2y^1|x}a = \lambda\Sigma_{y^1y^1|x}a, \quad (2)$$

$$\Sigma_{y^2y^1|x}\Sigma_{y^1y^1|x}^{-1}\Sigma_{y^1y^2|x}b = \lambda\Sigma_{y^2y^2|x}b. \quad (3)$$

By solving the problems (2) and (3), we get D (= $\max(d_{y^1}, d_{y^2})$) eigenvalues $\lambda_1 \geq ... \geq \lambda_d \geq ... \geq \lambda_D \geq 0$ and the corresponding eigenvectors a_1, ..., a_d, ..., a_D for y^1 and b_1, ..., b_d, ..., b_D for y^2. Here, the eigenvalues λ_d calculated from equations (2) and (3) are equivalent and equal to the square of the correlation coefficient ρ_d between $y^1|x$ and $y^2|x$. It can be proved that, assuming that the input data are taken from a Gaussian distribution, Partial CCA calculates the projections that maximize the TE (Shibuya, Harada, and Kuniyoshi 2011). Let us consider applying Partial CCA to two multivariate time series X and Y by substituting the past values of the time series x_{t-1} and the current values of the time series y_t for the input y^1 and y^2 respectively and the past value of the time series y_{t-1} for the third variable x. Then, by using eigenvector a_d, Y is projected to the subspace $S_d = a_d^T Y$ such that the TE from X to Y is the summation of the TE from X to S_d:

$$TE_{X\to Y} = \sum_{d=1}^{D} TE_{X\to S_d} = \frac{1}{2}\sum_{d=1}^{D}\log_2\frac{1}{1-\lambda_d}. \quad (4)$$

The causal effect $TE_{X\to S_d}$ is stronger when the index d is smaller and the corresponding eigenvalue λ_d is larger.

Thus, Partial CCA can analyze the causal effect in detail and allow an estimate of the causality measure robustly from multivariate data by projecting the data onto the subspace where the causal effect is maximized. Partial CCA is used in various fields, including bioinformatics (Fujita et al. 2010).

Bayesian Partial CCA

Partial CCA overfits the data when the number of samples exceeds the dimensionality of the data, because the covariance matrices become ill-conditioned. Group Sparse Bayesian Partial CCA (GSPCCA) (Mukuta and Harada

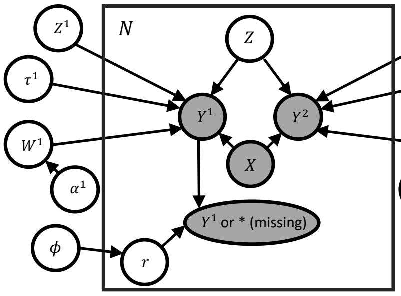


Figure 1: Graphical model of semi-BPCCA when only Y^1 can have missing values.

2014) prevents such overfitting by introducing a prior distribution to the model parameters.

GSPCCA assumes the generative model whereby two observed variables y^m $(m = 1, 2)$ are generated from a third observed variable x and the latent variable z is written as

$$\begin{aligned} z_n &\sim \mathcal{N}(0, I_{d_z+d_1+d_2}), \\ y_n^m |\, x_n,\, z_n &\sim \mathcal{N}(W_x^m x_n + W_z^m z_n, (\tau^m)^{-1} I_{d_m}). \end{aligned}$$

The prior distributions of the model parameters are

$$\begin{aligned} \alpha_k^m &\sim \text{Gamma}(a_0, b_0), \\ W_{:,k}^m &\sim \mathcal{N}(0, (\alpha_k^m)^{-1} I_{d_m}), \\ \tau^m &\sim \text{Gamma}(a_0, b_0), \end{aligned}$$

where Gamma(a, b) denotes the Gamma distribution with shape parameter a and scale parameter b; x_n, y_n^m, and z_n are n-th variables; and X, Y^m, and Z are the corresponding data matrices. Also, we denote Y=$({Y^1}^T {Y^2}^T)^T$ and W^m=$(W_x^m\ \ W_z^m)$.

GSPCCA yields unsatisfactory results when the number of complete samples is too small. In such a case, another constraint must be applied to the model parameters by exploiting the information contained in non-paired samples.

Proposed Method

In this section we propose the method for estimating the causal relationship between two incomplete time series by exploiting the information contained in non-paired samples. To this end, we propose a novel semi-supervised model called semi-Bayesian Partial CCA (semi-BPCCA), which is an extension of GSPCCA. Semi-BPCCA accepts non-paired samples as input and stabilizes the parameter estimation by incorporating the missingness of samples as a probabilistic variable into the model. It is similar to the approach of (Kamada, Kanezaki, and Harada 2015), where they extend probabilistic CCA to accept non-paired samples. We demonstrate that the bias of missing variables can be ignored without modeling the missing probability when the missing model is MCAR or MAR.

Semi-Bayesian Partial CCA

We constructed our model by considering Rubin's missing model (Rubin 1976). While our model can handle the case where both y^1 and y^2 have missing values, for simplicity, we here assume that only y^1 can have missing values. We introduce the random variable r_n as having a value of 1 when y_n^1 is missing and 0 otherwise. We model the probability of r_n as being controlled by the parameter ϕ. We represent a sample as $\{x_n, *, y_n^1, y_n^2\}$ where $*$ denotes the missing value. The graphical model that considers the missing values is plotted in Fig. 1. Denoting the model parameter $\Theta = \{Z, W, \tau_1, \tau_2, \alpha_1, \alpha_2\}$, the proposed method approximates the posterior probability of the parameter $p(\Theta|X, Y)$ as $q(\Theta)$ in a way that maximizes the data probability $p(X, Y)$ using the variational Bayesian method. The model evidence for the sample $\{x_n, *, y_n^2\}$ is decomposed as

$$\begin{aligned} L_n(q(\Theta), \phi) &= \log p(x_n, *, y_n^2, r_n|\phi) \qquad (5) \\ &= \log p(x_n, y_n^2) + \log \int p(y_n^1|x_n, y_n^2) p(r_n|x_n, y_n^1, y_n^2, \phi) dy_n^1. \end{aligned}$$

In the general case, one needs to model the probability $p(r_n|x_n, y_n^1, y_n^2, \phi)$ that the variable is missing to calculate the approximate posterior distribution $q(\Theta)$ that maximizes Equation (5). However, $q(\Theta)$ can be calculated without defining $p(r_n|x_n, y_n^1, y_n^2, \phi)$ for the specific missing distribution. Consider the three types of missing distributions Rubin defined. The first is MCAR, where the missing probability does not depend on y^1, y^2, or x and the distribution of r_t is independent of the y values. The second is MAR, where the missing probability depends on y^2 and x but does not depend on y^1 and the distribution of r_n is independent of y_n^1. The third is MNAR, where the missing probability depends on y^1 and the distribution of r_n depends on y_n^1. Considering the above three cases, when the missing distribution is MCAR or MAR, we can rewrite the second term in Equation (5) as the logarithm of $p(r_n|x_n, y_n^2, \phi)$. As this term is independent of Θ, the evidence, expressed in Equation (5), can be decomposed into terms that depend only on Θ and those depend only on ϕ as

$$\begin{aligned} L_n(q(\Theta), \phi) &= \log p(x_n, y_n^2) + \log p(r_n|x_n, y_n^2, \phi) \\ &= L_n(q(\Theta)) + L_n(\phi). \end{aligned}$$

Since the goal is to estimate $q(\Theta)$ when Θ and $L_n(\phi)$ are independent, $L_n(q(\Theta), \phi)$ can be maximized by maximizing $L_n(q(\Theta))$, independently of the missing model $L_n(\phi)$.

When all of the input data are segregated into complete samples, those samples where y^1 is missing, and those samples where y^2 is missing, the evidence of all the samples is written as

$$L(q(\Theta), \phi) = L^1(q(\Theta)) + L^2(q(\Theta)) + L^3(q(\Theta)) + const. \qquad (6)$$

where

$$\begin{aligned} L^m(q(\Theta)) &= \sum_{n \in S_m} L_n(q(\Theta)) = \sum_{n \in S_m} \log p(x_n, y_n^m), \\ L^3(q(\Theta)) &= \sum_{n \in S_3} L_n(q(\Theta)) = \sum_{n \in S_3} \log p(x_n, y_n^1, y_n^2). \end{aligned}$$

We write the terms that do not depend on Θ as "const" in Equation (6). S_m (m = 1, 2) denotes the set of samples in which only y^m is observed, and S_3 is the set of samples where both y^1 and y^2 are observed. The samples where the third variable x is missing are omitted. Estimating the Θ that maximize $L(q(\Theta, \phi))$ requires only the maximization of $L^1(q(\Theta)) + L^2(q(\Theta)) + L^3(q(\Theta))$. These terms are further rewritten as $\log p(X, {Y^1}^{S_1 \cup S_3}, {Y^2}^{S_2 \cup S_3})$, where Y^S is the matrix composed of $[y_n^1, y_n^2], (n \in S)$.

We calculate the distribution of Θ by variational Bayesian methods. We approximate the posterior $p(\Theta | X, {Y^1}^{S_1 \cup S_3}, {Y^2}^{S_2 \cup S_3})$ using $q(\Theta)$ written as

$$q(\Theta) = q(Z^{S_1}) q(Z^{S_2}) q(Z^{S_3}) \prod_m q(\tau^m) q(\alpha^m) q(W^m).$$

The variational approximate distributions of $\Theta = \{Z, W, \tau_1, \tau_2, \alpha_1, \alpha_2\}$ satisfy

$$q(\theta_i) \propto \exp\langle \log p(Y^{S_1 \cup S_3}, Y^{S_2 \cup S_3}, \Theta | X) \rangle_{j \neq i}, \quad (7)$$

where $\langle \bullet \rangle_{i \neq j}$ denotes the expectation with respect to $\prod_{j \neq i} q(\theta_j)$. We denote the approximate posterior of model parameters and latent variables as

$$\begin{aligned}
q(Z^{S_{m'}}) &= \prod_{n \in S_{m'}} \mathcal{N}(\mu_{z_n}^{S_{m'}}, \Sigma_z^{S_{m'}}) (m' = 1, 2, 3), \\
q(W^m) &= \prod_d \mathcal{N}(\mu_{W_{d,:}^m}, \Sigma_{W^m}), \\
q(\alpha^m) &= \prod_k \mathrm{Gamma}(a_{\alpha_k^m}, b_{\alpha_m}), \\
q(\tau^m) &= \mathrm{Gamma}(a_{\tau_m}, b_{\tau_m}),
\end{aligned}$$

and the distribution parameters are updated as

$$\begin{aligned}
&\Sigma_z^{S_m} = (I + \langle \tau^m \rangle \langle (W_z^m)^T (W_z^m) \rangle)^{-1}, \\
&\Sigma_z^{S_3} = (I + \langle \tau^1 \rangle \langle (W_z^1)^T (W_z^1) \rangle + \langle \tau^2 \rangle \langle (W_z^2)^T (W_z^2) \rangle)^{-1}, \\
&\langle Z \rangle^{S_m} = \Sigma_z^{S_m} \langle \tau^m \rangle (\langle W_z^m \rangle^T Y^{m S_m} - \langle (W_z^m)^T W_x^m \rangle X^{S_m}), \\
&\langle Z \rangle^{S_3} = \Sigma_z^{S_3} \big(\langle \tau^1 \rangle \big(\langle W_z^1 \rangle^T Y^{1 S_3} - \langle (W_z^1)^T W_x^1 \rangle X^{S_3} \big) \\
&\quad + \langle \tau^2 \rangle \big(\langle W_z^2 \rangle^T Y^{2 S_3} - \langle (W_z^2)^T W_x^2 \rangle X^{S_3} \big) \big), \\
&\Sigma_{W^m} = \left(\mathrm{diag}\langle \alpha^m \rangle + \langle \tau^m \rangle \begin{pmatrix} X^{S_3} (X^{S_3})^T & X^{S_3} (\langle Z \rangle^{S_3})^T \\ \langle Z \rangle^{S_3} (X^{S_3})^T & \langle Z Z^T \rangle^{S_3} \end{pmatrix} \right. \\
&\quad \left. + \langle \tau^m \rangle \begin{pmatrix} X^{S_m} (X^{S_m})^T & X^{S_m} (\langle Z \rangle^{S_m})^T \\ \langle Z \rangle^{S_m} (X^{S_m})^T & \langle Z Z^T \rangle^{S_m} \end{pmatrix} \right)^{-1}, \\
&\mu_{W^m} = Y^m ((X^{S_m \cup S_3})^T \langle Z^T \rangle^{S_m \cup S_3}) \Sigma_{W^m}, \\
&a_{\alpha^m} = a_0 + d_m / 2, \\
&b_{\alpha_k^m} = b_0 + \langle (W^m)^T (W^m) \rangle_{k,k} / 2, \\
&a_{\tau^m} = a_0 + (N_{S_m} + N_{S_3}) d_m / 2, \\
&b_{\tau^m} = b_0 + 0.5 \Big(\mathrm{Tr}\big(Y^m (Y^m)^T \big) \\
&\quad - 2\mathrm{Tr}\big(Y^m \big((X^{S_m \cup S_3})^T \langle Z^T \rangle^{S_m \cup S_3} \big) \langle W^m \rangle^T \big) \\
&\quad + \mathrm{Tr}\Big(\langle (W_z^m)^T (W_z^m) \rangle \begin{pmatrix} X^{S_3} (X^{S_3})^T & X^{S_3} (\langle Z \rangle^{S_3})^T \\ \langle Z \rangle^{S_3} (X^{S_3})^T & \langle Z Z^T \rangle^{S_3} \end{pmatrix} \\
&\quad + \langle (W_z^m)^T (W_z^m) \rangle \begin{pmatrix} X^{S_m} (X^{S_m})^T & X^{S_m} (\langle Z \rangle^{S_m})^T \\ \langle Z \rangle^{S_m} (X^{S_m})^T & \langle Z Z^T \rangle^{S_m} \end{pmatrix} \Big) \Big).
\end{aligned}$$

One advantage of this method is that it can stabilize the estimation of the covariance matrix Σ_W by using non-paired samples. In GSPCCA, the inverse of Σ_W is updated as

$$\Sigma_{W^m}^{-1} = \mathrm{diag}\langle \alpha^m \rangle + \langle \tau^m \rangle \begin{pmatrix} X^{S_3} (X^{S_3})^T & X^{S_3} (\langle Z \rangle^{S_3})^T \\ \langle Z \rangle^{S_3} (X^{S_3})^T & \langle Z Z^T \rangle^{S_3} \end{pmatrix}.$$

When the paired-sample size T_{pair} is small, the calculation of $\Sigma_{W^m}^{-1}$ becomes unstable. Especially when $T_{pair} > d = d_x + d_z$, the second term of $\Sigma_{W^m}^{-1}$ becomes ill-conditioned.

In semi-BPCCA, the reciprocal of Σ_W is updated as

$$\begin{aligned}
\Sigma_{W^m}^{-1} &= \mathrm{diag}\langle \alpha^m \rangle + \langle \tau^m \rangle \begin{pmatrix} X^{S_3} (X^{S_3})^T & X^{S_3} (\langle Z \rangle^{S_3})^T \\ \langle Z \rangle^{S_3} (X^{S_3})^T & \langle Z Z^T \rangle^{S_3} \end{pmatrix} \\
&\underbrace{+ \langle \tau^m \rangle \begin{pmatrix} X^{S_m} (X^{S_m})^T & X^{S_m} (\langle Z \rangle^{S_m})^T \\ \langle Z \rangle^{S_m} (X^{S_m})^T & \langle Z Z^T \rangle^{S_m} \end{pmatrix}}_{\text{non-paired}}. \quad (8)
\end{aligned}$$

By using non-paired samples, the third term of equation (8) serves as a regularization term to stabilize the estimation of Σ_{W^m}. In particular, when $T_{pair} < d$, this term prevents $\Sigma_{W^m}^{-1}$ from becoming ill-conditioned.

Also, the optimal W_z and Z have the degree of freedom. Thus, we optimize these W_z and Z as in GSPCCA to accelerate convergence. These distributions are optimized by setting $W_z^* = W_z R$ and $Z^* = R^{-1} Z$ and optimizing with respect to R. The objective function is defined as

$$\begin{aligned}
L(R) = &- \sum_{i=1}^{3} \frac{\mathrm{Tr}(R^{-1} \langle Z Z^T \rangle^{S_m} R^{-T})}{2} + (d_1 + d_2 - N) \mathrm{log}|R| \\
&- \frac{1}{2} \sum_{m=1}^{2} d_m \sum_{k=1}^{d_z} \log({r_k}^T \langle W_z^m {W_z^m}^T \rangle r_k).
\end{aligned}$$

R is optimized using the L-BFGS (Liu and Nocedal 1989) method after each update of model parameters, and the posterior distributions of W_z and Z are set as $\mu_{W_z^m} \leftarrow \mu_{W_z^m} R$, $\Sigma_{W_z^m} \leftarrow R^T \Sigma_{W_z^m} R$, $\langle Z \rangle \leftarrow R^{-1} \langle Z \rangle$, and $\Sigma_z \leftarrow R^{-1} \Sigma_z R^{-T}$.

Calculation of the Causality Measure

Semi-BPCCA can be applied to calculate the causality measure from time series $X = [x_1, x_2, ..., x_T]$ to $Y = [y_1, y_2, ..., y_T]$ by setting $Y^1 = X_{t-1}$, $Y^2 = Y_t$, and the third variable $X = Y_{t-1}$. In this work, we assume the embedding dimension to $k = l = 1$. The causal direction between X and Y is estimated by computing and comparing the causality measures from X to Y and from Y to X. When the ground-truth causal direction is $X \rightarrow Y$, it is desirable that $TE_{X \rightarrow Y} > TE_{Y \rightarrow X}$. We calculate the causality measure as $\sum_{d=1}^{d_z} \frac{1}{2} \log_2 \frac{1}{1 - \rho_d^2}$, where ρ_d is the correlation between the conditional expectations of the d-th dimension latent variable Z, calculated as

$$\begin{aligned}
\langle Z_{(d,:)} | X_{t-1}^{S_3}, Y_{t-1}^{S_3} \rangle &= \Sigma_z^{S_3} \langle \tau^2 \rangle (\langle W_z^2 \rangle^T X_{t-1}^{S_3} - \langle (W_z^2)^T W_x^2 \rangle Y_{t-1}^{S_3}), \\
\langle Z_{(d,:)} | Y_t^{S_3}, Y_{t-1}^{S_3} \rangle &= \Sigma_z^{S_3} \langle \tau^1 \rangle (\langle W_z^1 \rangle^T Y_t^{S_3} - \langle (W_z^1)^T W_x^1 \rangle Y_{t-1}^{S_3}).
\end{aligned}$$

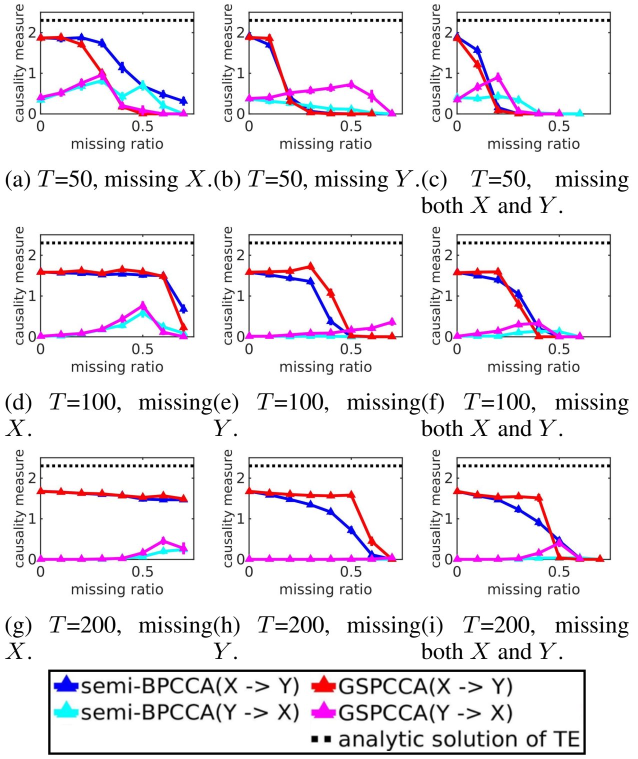
Only complete samples are used to compute the correlation.

Experiment on Synthesized Data

This section appraises the estimation of the causality measure using synthesized data. We compare the proposed semi-BPCCA and GSPCCA that ignore the non-paired samples from the input data.

Experimental Setting

Multi-dimensional data $X = [x_1, x_2, ..., x_T] \in R^{d_x \times T}$ and $Y = [y_1, y_2, ..., y_T] \in R^{d_y \times T}$ were generated such that



(a) T=50, missing X. (b) T=50, missing Y. (c) T=50, missing both X and Y.

(d) T=100, missing X. (e) T=100, missing Y. (f) T=100, missing both X and Y.

(g) T=200, missing X. (h) T=200, missing Y. (i) T=200, missing both X and Y.

Figure 2: Comparison of the estimated causality measures. The x- and y-axes correspond to the missing ratio and the causality measure, respectively. (MCAR)

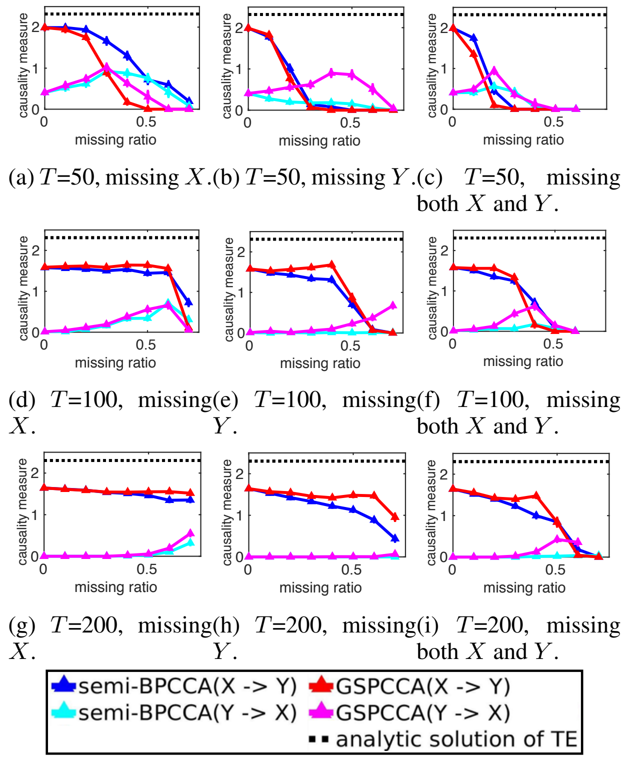
there is a causal direction from X to Y. Thus, it is desirable that the estimated $TE_{X\rightarrow Y}$ be greater than $TE_{Y\rightarrow X}$. The data were sampled using the following model:

$$\begin{aligned} x_t &= 0.5x_{t-1} + \epsilon_{t,x}, \\ y_{2_t} &= 0.5y_{2_{t-1}} + Wx_{t-1} + \epsilon_{t,y_2}, \\ y_t &= (y^T_{2_{t-1}}, y^T_{2_{t-1}})^T + \epsilon_{t,y}, \end{aligned} \tag{9}$$

where $\epsilon_{t,x}$, $\epsilon_{t,y_2} \sim \mathcal{N}(0, I)$ and $\epsilon_{t,y} \sim \mathcal{N}(0, 0.1I)$. W is a matrix where the entire first two columns were sampled from $\mathcal{N}(0, 0.5I)$ and the remaining values are 0. We can evaluate the estimation of Transfer Entropy by analytic solution induced from the model (9) and the definition of Transfer Entropy. Here, the value of the analytic solution of $TE_{X\rightarrow Y}$ is approximately 2.3, which is calculated as follows. Since x_t and y_t follow Gaussian distributions in this experiment, the analytic solution of $TE_{X\rightarrow Y}$ is

$$\begin{aligned} TE_{X\rightarrow Y} = &-\frac{1}{2}\log_2 \left|\Sigma_{\{y_t, y^{(l)}_{t-1}, x^{(k)}_{t-1}\}\{y_t, y^{(l)}_{t-1}, x^{(k)}_{t-1}\}}\right| \left|\Sigma_{y^{(l)}_{t-1}, y^{(l)}_{t-1}}\right| \\ &+ \frac{1}{2}\log_2 \left|\Sigma_{\{y_t, y^{(l)}_{t-1}\}\{y_t, y^{(l)}_{t-1}\}}\right| \left|\Sigma_{\{y^{(l)}_{t-1}, x^{(k)}_{t-1}\}\{y^{(l)}_{t-1}, x^{(k)}_{t-1}\}}\right| . \end{aligned} \tag{10}$$

Each covariance in Equation (10) can be computed from the covariance of the joint probability distribution $p(y_{t-1}, y_t, x_{t-1}, y_{2_{t-1}}, y_{2_t})$, which depends on the randomly sampled W. Meanwhile, the estimation of $TE_{Y\rightarrow X}$ is expected to become 0. We then removed the variable from



(a) T=50, missing X. (b) T=50, missing Y. (c) T=50, missing both X and Y.

(d) T=100, missing X. (e) T=100, missing Y. (f) T=100, missing both X and Y.

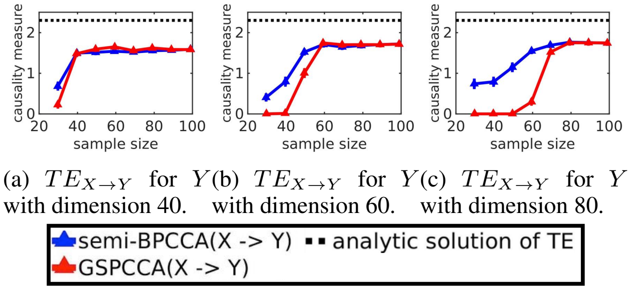
(g) T=200, missing X. (h) T=200, missing Y. (i) T=200, missing both X and Y.

Figure 3: Comparison of the estimated causality measures. The x- and y-axes correspond to the missing ratio and the causality measure, respectively. (MNAR)

x_t and y_t with the missing ratio $r = 0.1, 0.2, \ldots, 0.7$, and for the missing patterns of MCAR and MNAR. For MCAR, the variables are randomly removed uniformly. For MNAR, following (Kimura et al. 2010), we generated data missing X_t by removing x_t with the largest $r \times 100\%$ value of $a^T x_t$, where $a \sim N(0, I_{d_x})$. In the same way, data missing Y_t was generated by removing y_t with the largest $r \times 100\%$ value of $a^T y_t$, where $a \sim N(0, I_{d_y})$. We conducted the experiment for the case where only X_t was missing, only Y_t was missing, or both X_t and Y_t were missing. A missing variable in the outcome variable Y_t resulted in a missing value in the third variable Y_{t-1}. Thus, the ratio of complete pairs of $\{x_{t-1}, y_{t-1}, y_t\}$ is smaller than $1 - r$. We set the dimension of the latent variable z to 5. For each setting, we randomly generated the input data and calculated the causality measure 50 times and evaluated the mean and standard error of the estimated $TE_{X\rightarrow Y}$ and $TE_{Y\rightarrow X}$. If the estimated TE diverged in the 50 times trial, we set the TE of the trial to 0.

Effect of Missing Ratio

Fig. 2 and Fig. 3 show the results corresponding to the dimensionality of x_t being set to 20, that of y_t to 40, and the length of the time series T=50, 100, and 200 in the case of MCAR and MNAR respectively. The results of MNAR were similar to those of MCAR. This indicates our method can stabilize the calculation of causality measure regardless of the pattern of data missing. When T = 50, the proposed



(a) $TE_{X \to Y}$ for Y with dimension 40. (b) $TE_{X \to Y}$ for Y with dimension 60. (c) $TE_{X \to Y}$ for Y with dimension 80.

Figure 4: Comparison of the estimated causality measures with varying dimension. The x-axis is the size of paired samples after removing some of the samples from 100 samples.

method outperforms GSPCCA as it prevents the decrease of the causality measure caused by the shortage of paired samples even if the missing ratio becomes large. This result indicates the effectiveness of compensating the shortage by the information of non-paired samples. Also, our method is especially effective for estimating $TE_{X \to Y}$ when X has missing values and for estimating $TE_{Y \to X}$ when Y has missing values. Thus, our method exploits only those samples in which the outcome variable is observed. Our method is comparable to the GSPCCA in the case of T=200. Our method is also effective when there are few missing samples. For instance, Fig. 2 (e) illustrates that the causality measure estimated by GSPCCA approaches the analytic solution despite the missing ratio increasing from 0 to 0.4. This indicates that GSPCCA overfits the data with a small sample size and detects spurious causality. In contrast, our method prevents overfitting, as analyzed in detail in the following experiment. These results show that the proposed method is especially effective when the number of samples is small and the outcome variable is observed.

Effect of the Dimension of the Data

We here evaluate the effect of the dimensionality of the input data. We set the sample size $T = 100$, the dimensionality of x_t to 20, and vary the dimensionality of y_t to 40, 60, or 80.

Fig. 4 shows that our method is effective when the number of the samples is smaller than the dimensionality of y_t.

Analysis of the Eigenvalues

We then decomposed $TE_{X \to Y}$ to the TE corresponding to each subspace of the 5 greatest eigenvalues for detailed analysis. Here, the projection of a variable onto the subspace associated with a greater eigenvalue has a stronger causal relationship with the paired variable.

Fig. 5 shows that, while the third and fourth eigenvalues estimated by GSPCCA become large as the missing ratio increases, the corresponding values estimated by our methods are relatively small and the fourth eigenvalue is 0. The synthesized data we used did not have a causal effect for the direction corresponding to the third and fourth eigenvectors, but the GSPCCA overfits the data and detects the wrong causal relationship. As for the first eigenvalue, while the causality measure increases as the missing value increases for GSPCCA, our method yields a stable value. Thus, our method prevents from detecting the wrong causality by exploiting the non-paired samples. Compared to the GSPCCA, the proposed method works well when $TE_{X \to Y}$ is calculated using the data where only Y is observed.

Experiments on Real Data

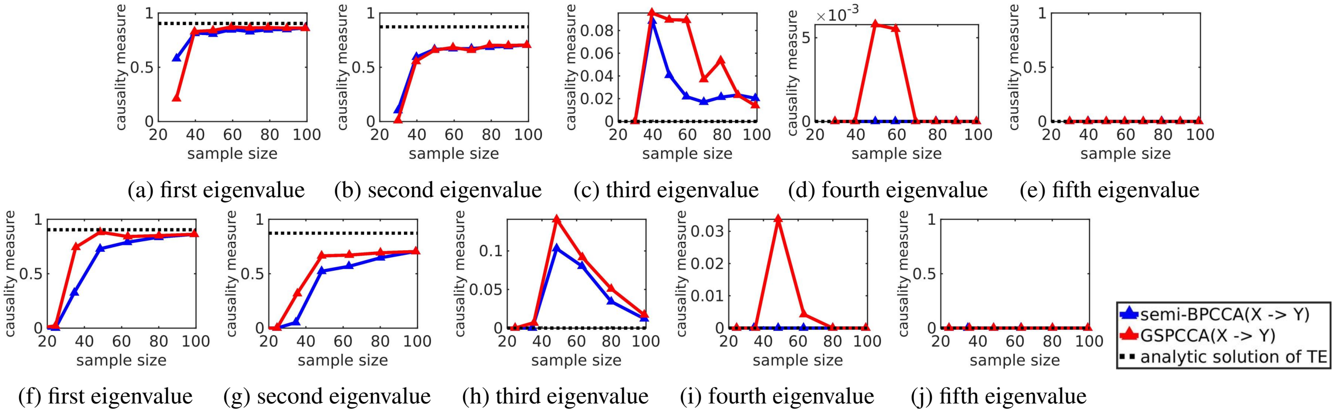
This section evaluates our method on meteorological data and video analysis datasets. Both experiments assume that the missing distribution is MCAR and that both the cause and outcome variables have missing values.

Experiment on Meteorological Data

In meteorological studies, atmosphere dynamics are analyzed by simulations. However, it is difficult to estimate a large number of parameters in a complex model. In contrast, we assume that our method can approximately predict the weather transition as information flow. We tested its performance using the meteorological data with missing values.

We evaluated the performance of the estimation of climatic information flow using the Global Summary of the Day, the meteorological dataset that contains information about climatic element observed by the National Climatic Data Center. Fig. 6 (a) shows the jet stream over the North American continent. There is interest in quantifying the global flow of the atmosphere using data of local climate elements collected in cities. We chose data obtained from 224 stations located all over the USA. We selected seven types of climate elements that have few missing values: the mean temperature, maximum temperature, minimum temperature, mean wind speed, maximum sustained wind speed, mean dew point, and mean visibility. We used data taken during the winter season spanning 67 days from Dec. 24, 2008 to Feb. 28, 2009. To control the missing ratio, we first did a zero-order hold for the missing values in the dataset. After that, we randomly remove 20% or 40% of the sample size from the data taken at each station. After calculating the causality measures between all pairs of stations and eliminating diverging results, we visualized the 50 largest information flows. We set the dimensionality of the latent variable to 10. When the missing ratio is 0%, the models of our method and GSPCCA are equivalent.

Figures 6 (b-f) display the results. When the missing ratio is 20%, our method outperformed GSPCCA in quantifying the flow from north to south in the central region. When the missing ratio is 40%, some of the flows quantified by GSPCCA are opposite to the direction of the actual air mass movement seen in Fig. 6 (a). On the other hand, our method successfully quantifies the flow from west to east in the eastern region and the flow from northwest to southeast. These results estimated by our results are consistent with the actual jet stream, shown in Fig. 6 (a). Next, we calculated the average arrow length (km/day) using the Hubeny formula (http://www.kashmir3d.com/kash/manual-e/std_siki.htm). Table 1 shows that the proposed method is more stable than GSPCCA when the missing ratio is large. It is also similar to the result calculated when the missing ratio is zero. These results show that our method is



(a) first eigenvalue (b) second eigenvalue (c) third eigenvalue (d) fourth eigenvalue (e) fifth eigenvalue

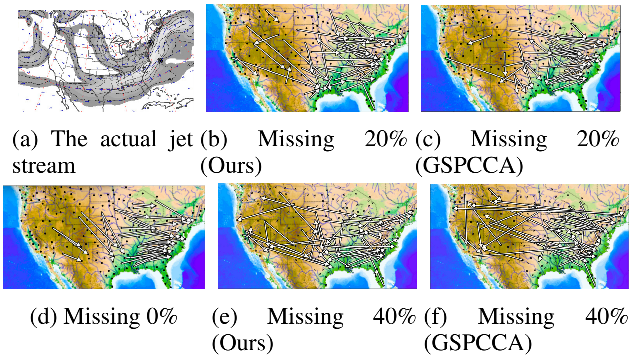
(f) first eigenvalue (g) second eigenvalue (h) third eigenvalue (i) fourth eigenvalue (j) fifth eigenvalue

Figure 5: Comparison of estimated $TE_{X \to S_k}$ corresponding to the 5 greatest eigenvalues. Cases (a-e) were estimated with an incomplete X and cases (f-j) with incomplete Y. The x-axis represents the size of paired samples after removing some of the samples from 100 samples.

effective in providing an approximate prediction for the transition of the atmosphere without using physics-based models and quantifying the strength of the information flow even when the data have missing values.

Experiment on Causal Flow

Next, we apply semi-BPCCA (our method) and GSPCCA to video datasets with missing pixel values to detect the global pattern of motion within a scene. Optical flow is a well-known method for motion detection in a video stream. Object motion is expressed using a vector that represents the relative displacement of pixels between frames. Causal flow (Yamashita, Harada, and Kuniyoshi 2012) is another approach that regards the object motion in a video as information flow between neighboring pixels, i.e., reflecting the effect of one pixel value on the value of a neighboring pixel. We can employ the causality measure to quantify such causal effect. The causal flow approach assumes that none of the pixels are missing. However, occasional outlier pixels count as effectively missing values. This experiment follows the causal flow approach and demonstrates that our method performs well even in the case of a substantial number of missing pixels in the video frames. To calculate the causal flow, we constructed time series from pixel values and calculated the intensity of causal effect from one pixel to its eight neighboring pixels. Then, we integrated the intensities in the eight directions to quantify the orientation and the intensity of the information flow from the pixel. In this experiment, we used the feature extracted from each frame using a pre-trained convolutional neural network instead of pixel values to extract high-dimensional semantic information. We first resized the input frame to 224×224 and then extracted the output of the second pooling layer of VGG-16 (Simonyan and Zisserman 2014). Thus, the dimensionality of the time series is 128 and the length of the time series is the number of video frames. We calculate causality measure between pixels using our method and GSPCCA. For each pixel of the video clip of T frames, we randomly removed $T \times r$ pixel values uniformly. We used the Crowd Segmentation Dataset (Ali and Shah 2007), which contains video clips of crowded scenes and is available on the Internet.



(a) The actual jet stream (b) Missing 20% (Ours) (c) Missing 20% (GSPCCA)

(d) Missing 0% (e) Missing 40% (Ours) (f) Missing 40% (GSPCCA)

Figure 6: (a): The jet stream observed in February 2009. (b-f) Climatic information flow map estimated from the data for the USA in the winter of 2009.

Table 1: Comparison of the average arrow lengths.

Missing Ratio	0%	20%	40%
GSPCCA	9.89×10^2	1.01×10^3	1.28×10^3
Ours	9.89×10^2	1.04×10^3	1.09×10^3

The results are shown in Fig. 7. The left, central, and right columns are the results of the videos with 325, 327 and 202 frames, respectively. Overall, when the missing ratio is 30%, our method successfully captures the motion flow similarly to the result when the missing ratio is 0%, whereas GSPCCA cannot capture any motion flow. For video C in Fig. 7, it is especially difficult to capture motion because the number of frames is so small that a missing ratio of 30% amounts to 140 frames with no missing pixels, which nearly equals 128, the dimensionality of the feature. Under such an extreme condition, our method detects the motion in the middle region of the frames. On the other hand, GSPCCA captures spurious flow in the upper regions of the frames.

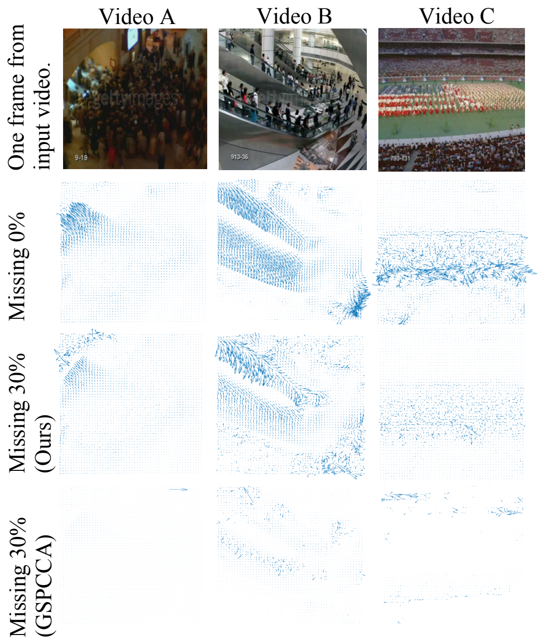


Figure 7: Estimated causal flow from the original video and from the video with missing values.

Conclusion

The present study has been the first to address the problem of estimating the causality measure between incomplete time series. We proposed a semi-supervised extension of Bayesian Partial CCA, called semi-Bayesian Partial CCA, that can exploit the information of samples with missing values. In our model, those samples effectively regularize the covariance matrix necessary to estimate the model parameters that are significant in quantifying causal relationships. We also demonstrated that our method can estimate the causal effect independently of the probabilistic model used for missing data, provided that the missingness is independent of the missing values. Experimental results based on artificial data confirmed the usefulness of the proposed method when there are few samples and the outcome variable is observed. Experiments on real-world data also demonstrated that the proposed method performs well when the number of complete samples is so small that existing methods overfits and finds spurious causal relationships. Future extensions of our method can consider, e.g., the modeling of nonlinear dynamics or of a delay between cause and effect, or model variables that do not follow the Gaussian distribution such as count data using an exponential family distribution.

Acknowledgments

The work was partially supported by JST CREST Grant Number JPMJCR1403, Japan and was partially funded by the ImPACT Program of the Council for Science, Technology, and Innovation (Cabinet Office, Government of Japan).

References

Ali, S., and Shah, M. 2007. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *CVPR*.

Berger, E.; Grehl, S.; Vogt, D.; Jung, B.; and Amor, H. 2016. Experience-based torque estimation for an industrial robot. In *ICRA*.

Bishop, C. M. 1999. Bayesian pca. In *NIPS*, 382–388.

Fujita, A.; Sato, J. R.; Kojima, K.; Gomes, L. R.; Nagasaki, M.; Sogayar, M. C.; and Miyano, S. 2010. Identification of granger causality between gene sets. *Journal of Bioinformatics and Computational Biology* 8(04):679–701.

Granger, C. W. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37(3):424–438.

Ilin, A., and Raiko, T. 2010. Practical approaches to principal component analysis in the presence of missing values. *JMLR* 11(Jul):1957–2000.

Kamada, C.; Kanezaki, A.; and Harada, T. 2015. Probabilistic semi- canonical correlation analysis. In *ACMMM*.

Kimura, A.; Kameoka, H.; Sugiyama, M.; Nakano, T.; Sakano, E. M. H.; and Ishiguro, K. 2010. Semicca: Efficient semisupervised learning of canonical correlations. In *ICPR*.

Klami, A.; Virtanen, S.; and Kaski, S. 2013. Bayesian canonical correlation analysis. *JMLR* 14(Apr):965–1003.

Kwon, O., and Yang, J.-S. 2008. Information flow between stock indices. *Europhysics Letters* 82(6):68003.

Liu, D. C., and Nocedal, J. 1989. On the limited memory bfgs method for large scale optimization. *Mathematical programming*.

Mukuta, Y., and Harada, T. 2014. probabilistic partial canonical correlation analysis. In *ICML*.

Rao, B. R. 1969. Partial canonical correlations. *Trabajos de Estadistica y de Investigacion operative* 20(2):211–219.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66(5):688.

Rubin, D. B. 1975. Bayesian inference for causality: The importance of randomization. In *The Proceedings of the social statistics section of the American Statistical Association*, 233–239.

Rubin, D. B. 1976. Inference and missing data. *Biometrika* 63(3):581–592.

Schreiber, T. 2000. Measuring information transfer. *Physical Review Letters* 85(2):461.

Shibuya, T.; Harada, T.; and Kuniyoshi, Y. 2009. Causality quantification and its applications: structuring and modeling of multivariate time series. In *KDD*.

Shibuya, T.; Harada, T.; and Kuniyoshi, Y. 2011. Reliable index for measuring information flow. *Physical Review E* 84(6):061109.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. In *ICPR*.

Wibral, M.; Pampu, N.; Priesemann, V.; Siebenhühner, F.; Seiwert, H.; Lindner, M.; Lizier, J. T.; and Vicente, R. 2013. Measuring information-transfer delays. 8(2):e55809.

Yamashita, Y.; Harada, T.; and Kuniyoshi, Y. 2012. Causal flow. In *ICME*.