Probabilistic Partial Canonical Correlation Analysis

Yusuke Mukuta MUKUTA@MI.T.U-TOKYO.AC.JP
Graduate School of Information Science and Technology, The University of Tokyo
7–3–1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan

Tatsuya Harada HARADA@MI.T.U-TOKYO.AC.JP
Graduate School of Information Science and Technology, The University of Tokyo
7–3–1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan

Abstract

Partial canonical correlation analysis (partial CCA) is a statistical method that estimates a pair of linear projections onto a low dimensional space, where the correlation between two multidimensional variables is maximized after eliminating the influence of a third variable. Partial CCA is known to be closely related to a causality measure between two time series. However, partial CCA requires the inverses of covariance matrices, so the calculation is not stable. This is particularly the case for high-dimensional data or small sample sizes. Additionally, we cannot estimate the optimal dimension of the subspace in the model. In this paper, we have addressed these problems by proposing a probabilistic interpretation of partial CCA and deriving a Bayesian estimation method based on the probabilistic model. Our numerical experiments demonstrated that our methods can stably estimate the model parameters, even in high dimensions or when there are a small number of samples.

1. Introduction

Partial canonical correlation analysis (partial CCA) was proposed by Rao (1969). It is a statistical method used to estimate a pair of linear projections onto a low-dimensional space, where the correlation between two multidimensional variables is maximized after eliminating the influence of a third variable. This is calculated using a CCA of the residuals of a linear regression of the third variable. This method is a generalized version of the partial correlation coefficient for multidimensional data. We define the variables $\{y_n^1\}_{n=1}^N \in \mathbb{R}^{d_1}$ and $\{y_n^2\}_{n=1}^N \in \mathbb{R}^{d_2}$, the third variable $\{x_n\}_{n=1}^N \in \mathbb{R}^{d_x}$ and the dimension of the subspace d_z. Then the partial CCA is calculated using the general eigenvalue problem

$$\begin{aligned}\Sigma_{12|x}\Sigma_{22|x}^{-1}\Sigma_{21|x}u^1 &= \rho^2\Sigma_{11|x}u^1,\\ \Sigma_{21|x}\Sigma_{11|x}^{-1}\Sigma_{12|x}u^2 &= \rho^2\Sigma_{22|x}u^2, \qquad (1)\end{aligned}$$

where $\Sigma_{m_1m_2|x} = \Sigma_{m_1x} - \Sigma_{m_1x}\Sigma_{xx}^{-1}\Sigma_{xm_2}$, and Σ_{ab} is a sample covariance matrix. Partial CCA has various applications in areas such as social science (Kowalski et al., 2003), and can be used as a causality measure.

Causality measures are indices that measure the influence of one time series on another. Transfer entropy (Schreiber, 2000) is a measure based on information theory. It measures the magnitude of a change to the conditional distribution of y given x, and is calculated using

$$\begin{aligned}T_{x\to y} &= \iiint p(y_t, y_{t-1}^{(l)}, x_{t-1}^{(k)})\\ &\qquad \log_2 \frac{p(y_t|y_{t-1}^{(l)}, x_{t-1}^{(k)})}{p(y_t|y_{t-1}^{(l)})} dy_t dy_{t-1}^{(l)} dx_{t-1}^{(k)}, \qquad (2)\end{aligned}$$

where k and l denote the embedding dimensions, $y_{t-1}^{(l)} = \left(y_{t-1}^T\, y_{t-2}^T \cdots y_{t-l+1}^T\right)^T$, and $x_{t-1}^{(k)} = \left(x_{t-1}^T\, x_{t-2}^T \cdots x_{t-k+1}^T\right)^T$. Shibuya et al. (2009) showed that when we assume that the variables are normally distributed and estimate the model parameters using maximum likelihood estimation, transfer entropy is equivalent to Granger causality (Granger, 1969). Granger causality is based on changes to the estimation error of an autoregressive model. Shibuya et al. (2011) showed that we can use the partial canonical correlations, ρ_i, calculated using partial CCA on y_t and $x_{t-1}^{(k)}$ and eliminate the effect of $y_{t-1}^{(l)}$. Then, the transfer entropy can be calculated using $T_{x\to y} = \frac{1}{2}\sum_{i=1}^{\min(d_y, kd_x)} \log_2 \frac{1}{1-\rho_i^2}$. Transfer entropy

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

has many applications such as brain analysis (Chávez et al., 2003), medical science (Verdes, 2005), cognitive development modelling (Sumioka et al., 2008), and detecting motion in a movie (Yamashita et al., 2012).

However, partial CCA requires the inverses of sample covariance matrices, so the calculation is unstable when the variables are highly correlated, the dimension of the data is large, or there are not enough data. Yamashita et al. regularized the covariance matrix to solve this problem (2012), but the appropriate optimization of the plural regularization parameters has not been determined. Additionally, we cannot estimate the proper dimension of the subspace of the model.

We have addressed these problems by proposing a probabilistic interpretation of partial CCA, and by deriving a variational Bayesian estimation algorithm for the model parameters based on this probabilistic interpretation. Our experiments show that the proposed methods can more accurately estimate the subspace dimension, and can more stably estimate the model parameters on both synthetic and real data, even in high dimensions or when there are few samples.

2. Canonical Correlation Analysis and its Extension

In this section, we review canonical correlation analysis, which is a statistical method similar to partial CCA. We also consider it from a probabilistic perspective.

Canonical correlation analysis (CCA) was proposed by Hotelling (1936). It is a method for finding statistical dependencies between two data sources. Given variables $\{y^1_n\}_{n=1}^N \in \mathbb{R}^{d_1}$ and $\{y^2_n\}_{n=1}^N \in \mathbb{R}^{d_2}$, and the dimension of the subspace $d_z \leq \min(d_1, d_2)$, the CCA can be calculated using the general eigenvalue problem

$$\begin{aligned}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}u^1 &= \rho^2\Sigma_{11}u^1,\\ \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}u^2 &= \rho^2\Sigma_{22}u^2,\end{aligned} \tag{3}$$

where $\Sigma_{m_1 m_2}$ represents a sample covariance matrix between y^{m_1} and y^{m_2}. The projection is a $d_z \times d_i$ $(i = 1, 2)$ matrix with the d-th row eigenvector corresponding to the d-th largest eigenvalue. Each eigenvalue equals the correlation in each dimension. Numerous studies have extended CCA, including a nonlinear extension using kernels (Lai & Fyfe, 2000; Melzer et al., 2001), online inferences of the model parameters (Vía et al., 2007; Yger et al., 2012), and sparse variants (Hardoon & Shawe-Taylor, 2009).

Bach and Jordan gave a probabilistic interpretation of CCA (2005), such that the maximum likelihood estimates of the model parameters can be derived from the CCA. Given this probabilistic interpretation, we can extend CCA to probabilistic models. Figure 1 shows a graphical model of the interpretation, where $\{z_n\}_{n=1}^N \in \mathbb{R}^{d_z}$ are the latent variables. The generative model is

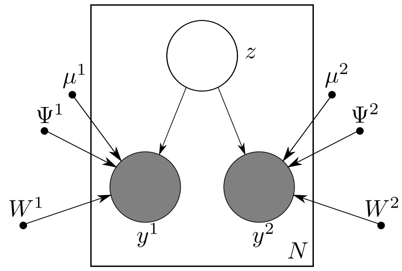


Figure 1. Graphical model for probabilistic CCA.

$$\begin{aligned} z_n &\sim \mathcal{N}(0, I_{d_z}),\\ y_n^m &\sim \mathcal{N}\left(W^m z_n + \mu^m, \Psi^m\right),\end{aligned} \tag{4}$$

where $\mathcal{N}(\mu, \Sigma)$ denotes the multivariate normal distribution with mean μ and covariance Σ, and I_d denotes the d dimensional identity matrix. $W^m \in \mathbb{R}^{d_m \times d_z}$ and $\Psi^m \in \mathbb{R}^{d_m \times d_m}$ are the model parameters that we must estimate. We define the $U^m_{d_z}$ matrices as having their d-th column equal to the d-th eigenvector, and $P_{d_z} \in \mathbb{R}^{d_m \times d_z}$ as a diagonal matrix with d-th element equal to the d-th eigenvalue of Equation (3). Then, the maximum likelihood solution is

$$\begin{aligned} W^m &= \Sigma_{mm}U^m_{d_z}M_m,\\ \Psi^m &= \Sigma_{mm} - W^m(W^m)^T,\\ \mu^m &= \overline{y}^m,\end{aligned} \tag{5}$$

where $M_m \in \mathbb{R}^{d_m \times d_m}$ are arbitrary matrices such that $M_1 M_2^T = P_{d_z}$ and the spectral norms of M_m are smaller than one. $\overline{y}^m$ is the sample mean $\frac{1}{N}\sum_{n=1}^N y_n^m$. There are some extensions of this probabilistic model. They include a robust estimation method that assumes a student distribution for noise (Archambeau et al., 2006), and a nonlinear extension that uses a Gaussian process latent variable model (Leen & Fyfe, 2006; Ek et al., 2008).

Bayesian CCA (Klami & Kaski, 2007; Wang, 2007) assumes that the model parameters are also random variables. Wang used a Wishart prior for the precision matrices of the noise, an ARD prior (Neal, 1995) for each column of the projection matrices, and derived a variational Bayesian estimation algorithm for the posterior distribution of the parameters. Virtanen et al. (2011) reduced the number of model parameters by assuming that the noise was isotropic and by introducing non-shared latent variables. Klami et al. (2013) derived an algorithm that simultaneously inferred the projection matrices for the shared and non-shared variables. Damianou et al. (2012) studied a Bayesian extension of a Gaussian process latent variable model. Fujiwara et al. (2009) used Bayesian CCA to estimate image bases from fMRI data.

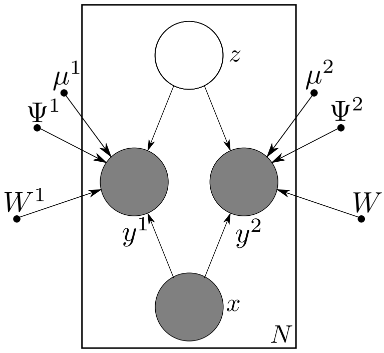


Figure 2. Graphical model for probabilistic partial CCA.

3. Probabilistic Interpretation of Partial CCA

In this section, we propose a generative model that estimates the maximum likelihood parameters using partial CCA. We also derive an expectation-maximization (EM) algorithm that estimates the model parameters and latent variables.

3.1. Generative Model

We consider a generative model that combines the regressions of variables that have effects we want to eliminate and shared latent variables, as shown in Figure 2. The model is defined as

$$\begin{array}{rcl} z_n & \sim & \mathcal{N}(0, I_{d_z}), \\ y_n^m & \sim & \mathcal{N}\left(W_x^m x_n + W_z^m z_n + \mu^m, \Psi^m\right). \end{array} \quad (6)$$

We will show that the maximum likelihood solution $\underset{W_x, W_z, \Psi}{\arg\max} \log p(y|x; W_x, W_z, \Psi)$ can be calculated using partial CCA. To this end, we show that the proposed model can be reduced to the generative model of probabilistic CCA Equation (4). When we define the log likelihood L and

$$C = \begin{pmatrix} \Psi^1 & 0 \\ 0 & \Psi^2 \end{pmatrix} + \begin{pmatrix} W_z^1 \\ W_z^2 \end{pmatrix} \begin{pmatrix} {W_z^1}^T {W_z^2}^T \end{pmatrix},$$

it holds that

$$\frac{\partial L}{\partial \mu} = -\sum_{n=1}^{N} C^{-1} \left(\begin{pmatrix} \mu^1 \\ \mu^2 \end{pmatrix} - \begin{pmatrix} y_n^1 \\ y_n^2 \end{pmatrix} + \begin{pmatrix} W_x^1 \\ W_x^2 \end{pmatrix} x_n \right).$$

Because C is positive definite, the likelihood is maximized when μ is such that the partial derivative equals zero. Therefore,

$$\mu^m = \overline{y}^m - W_x^m \overline{x}. \quad (7)$$

We denote each datum minus the sample mean as $\widetilde{y}_n^1 = y_n^1 - \overline{y}^1$, and substitute Equation (7). Then,

$$\frac{\partial L}{\partial W_x} = \sum_{n=1}^{N} C^{-1} \left(\begin{pmatrix} \widetilde{y}_n^1 \\ \widetilde{y}_n^2 \end{pmatrix} \widetilde{x}_n^T - \begin{pmatrix} W_x^1 \\ W_x^2 \end{pmatrix} \widetilde{x}_n \widetilde{x}_n^T \right).$$

We can also show that if the data space is spanned by the samples, L is the negative definite quadratic form of W_x. So L is maximized when W_x is such that the partial derivative is zero. Therefore,

$$W_x^m = \Sigma_{mx} \Sigma_{xx}^{-1}. \quad (8)$$

When we substitute this into Equation (6), the model is equivalent to the probabilistic CCA model with input variables ${y'}_n^m = \widetilde{y}_n^m - \Sigma_{mx}\Sigma_{xx}^{-1}\widetilde{x}_n$. Because the covariance matrices of these data are

$$\begin{array}{rcl} \dfrac{1}{N} \displaystyle\sum_{n=1}^{N} {y'}_n^{m_1} {y'}_n^{m_2 T} & = & \Sigma_{m_1 m_2} - \Sigma_{m_1 x} \Sigma_{xx}^{-1} \Sigma_{x m_2} \\ & = & \Sigma_{m_1 m_2 | x}, \end{array} \quad (9)$$

the parameter estimation is reduced to partial CCA. To summarize, the maximum likelihood solution of the proposed model can be written as

$$\begin{array}{rcl} W_x^m & = & \Sigma_{mx} \Sigma_{xx}^{-1}, \\ W_z^m & = & \Sigma_{mm|x} U_{d_z}^m M_m, \\ \Psi^m & = & \Sigma_{mm|x} - W_z^m W_z^{mT}, \\ \mu^m & = & \overline{y}^m - W_x^m \overline{x}, \end{array} \quad (10)$$

where $U_{d_z}^m$ denotes matrices that have their d-th column equal to the d-th eigenvector, P_d denotes the diagonal matrix with its d-th element equal to the d-th canonical correlation of Equation (1), and M_m are arbitrary matrices that satisfy $M_1 M_2^T = P_{d_z}$ and have spectral norms smaller than one. From this point, we assume that samples have zero mean and we do not infer a sample mean.

3.2. EM Parameter Estimation

As with CCA, we can estimate the latent variables using the EM algorithm without integrating them out. In this case, z_n follows a normal distribution and the update rule for time t is

$$\begin{array}{rcl} (\Sigma_z)_t & = & (I + (W_z)_t{}^T (\Psi_t)^{-1} (W_z)_t)^{-1}, \\ \langle Z \rangle_t & = & (\Sigma_z)_t (W_z)_t{}^T (\Psi_t)^{-1} (Y - (W_x)_t X), \\ W_{t+1}^m & = & Y^m \begin{pmatrix} X \\ \langle Z \rangle_t \end{pmatrix}^T \begin{pmatrix} X X^T & X \langle Z \rangle_t^T \\ \langle Z \rangle_t X^T & \langle Z Z^T \rangle_t \end{pmatrix}^{-1}, \quad (11) \\ \Psi_{t+1}^m & = & \dfrac{1}{N} \left(Y^m Y^{mT} - \left(W_{t+1} \begin{pmatrix} X \\ \langle Z \rangle_t \end{pmatrix} Y^T \right)_{mm} \right), \end{array}$$

where Ψ_t is the matrix with Ψ_t^m on its diagonal, W_x, and W_z are the matrices that have W_x^m and W_z^m in their columns, A_{mm} is the block matrix corresponding to each view, Y^m is the matrix that has y_n^m in its rows, and $Y = \begin{pmatrix} Y^1 \\ Y^2 \end{pmatrix}$. Additionally, X and Z are matrices with x_n and y_n in their rows, and $\langle \cdot \rangle$ are the expectations of the random variables.

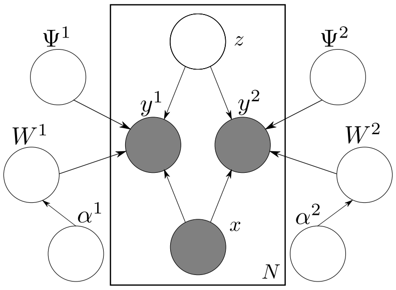


Figure 3. Graphical model for BPCCA.

4. Bayesian Partial CCA

To address the previously mentioned weakness of partial CCA, we propose a hierarchical Bayesian approach to the probabilistic partial CCA proposed in the previous section.

4.1. Model that Directly uses Probabilistic Partial CCA

In this section, we follow Wang's approach (2007) and consider the generative model shown in Figure 3. It treats the model parameters proposed in the previous section as random variables. We use an ARD prior (Neal, 1995) for each column of the projection matrices, and an inverse Wishart prior for the covariance matrices of the noise. The generative model is

$$\begin{aligned}
\alpha_k^m &\sim \mathrm{Gamma}(a_0, b_0),\\
W_{:,k}^m &\sim \mathcal{N}(0, (\alpha_k^m)^{-1} I_{d_m}),\\
\Psi^m &\sim \mathcal{IW}(\nu_0^m, K_0^m),\\
z_n &\sim \mathcal{N}(0, I_{d_z}),\\
y_n^m &\sim \mathcal{N}(W_x^m x_n + W_z^m z_n, \Psi^m),
\end{aligned} \tag{12}$$

where the prior for the third variable $p(x)$ does not affect the inference when $p(x_n) > 0$ for each sample, because we consider the conditional distribution given x_n. Here $\mathrm{Gamma}(a, b)$ is the Gamma distribution with shape parameter a and scale parameter b, and $\mathcal{IW}(\nu, K)$ is the inverse Wishart distribution. $W^m = \begin{pmatrix} W_x^m & W_z^m \end{pmatrix}$. $W_{:,k}^m$ is the k-th column of W^m. The hyperparameters a_0, b_0, ν_0^m, K_0^m should be small so that the priors are broad, but from the definition of the Wishart distribution, $\nu_0^m > d_m - 1$. In our experiments, we set $a_0, b_0 = 10^{-14}, \nu_0^m = d_m, K_0^m = 10^{-14} \cdot I_{d_m}$. The ARD prior drives unnecessary components to zero, so we can estimate the dimensions of the latent variables by choosing sufficiently large d_z, or by first choosing a small d_z and then gradually increasing it according to the output projection matrices. We refer to this model as Bayesian PCCA (BPCCA).

Next, we propose a variational Bayesian inference algorithm. The full posterior $p(Z, \Theta | X, Y)$ is approximated as

$$q(Z, \Theta) = q(Z) \prod_{m=1}^{2} \left(q(\Psi^m) q(\alpha^m) \prod_{j=1}^{d_m} q(w_j^m) \right), \tag{13}$$

where w_j^m is the j-th row of W^m. We apply standard cyclical updates to the separate terms of q. When the factorized distribution q has the form $\prod_i q(\theta_i)$, the update rule is

$$\begin{aligned}
q(\theta_i) &\propto \exp\left(\langle \log p(X, Y, Z, \Theta) \rangle_{Z, \theta_{k \neq i}} \right),\\
q(Z) &\propto \exp\left(\langle \log p(X, Y, Z, \Theta) \rangle_{\Theta} \right).
\end{aligned} \tag{14}$$

Because $p(X)$ is independent of the other variables, it follows that

$$\begin{aligned}
q(\theta_i) &\propto \exp\left(\langle \log p(Y, Z, \Theta | X) \rangle_{Z, \theta_{k \neq i}} \right),\\
q(Z) &\propto \exp\left(\langle \log p(Y, Z, \Theta | X) \rangle_{\Theta} \right),
\end{aligned} \tag{15}$$

where $\langle \cdot \rangle$ with subscripts denote the expectation with respect to the approximate posterior distribution of the corresponding variables. The approximate posterior distribution has the shape

$$\begin{aligned}
q(z_n) &= \mathcal{N}(\mu_{z_n}, \Sigma_{z_n}),\\
q(\Psi^m) &= \mathcal{IW}(\nu_m, K_m),\\
q(w_j^m) &= \mathcal{N}(\mu_{w_j^m}, \Sigma_{w_j^m}),\\
q(\alpha^m) &= \prod_k \mathrm{Gamma}(a_m, b_{mk}).
\end{aligned} \tag{16}$$

Furthermore, the parameters are updated as

$$\begin{aligned}
\Sigma_{z_n} &= \left(I + \sum_m \langle (W_z^m)^T (\Psi^m)^{-1} W_z^m \rangle \right)^{-1},\\
\mu_{z_n} &= \Sigma_{z_n} \sum_m \left(\langle (W_z^m)^T \rangle \langle (\Psi^m)^{-1} \rangle y_n^m \right.\\
&\quad \left. - \langle (W_z^m)^T (\Psi^m)^{-1} W_x^m \rangle x_n \right),\\
K_m &= K_0^m + Y^m (Y^m)^T\\
&\quad + \langle W^m \begin{pmatrix} XX^T & XZ^T \\ ZX^T & ZZ^T \end{pmatrix} (W^m)^T \rangle\\
&\quad - Y^m \begin{pmatrix} X^T & \langle Z^T \rangle \end{pmatrix} \langle (W^m)^T \rangle\\
&\quad - \langle W^m \rangle \begin{pmatrix} X \\ \langle Z \rangle \end{pmatrix} Y^m,\\
\nu_m &= \nu_0^m + N,\\
\Sigma_{w_j^m} &= \left(\mathrm{diag} \langle \alpha^m \rangle + \langle (\Psi^m)_{j,j}^{-1} \rangle \begin{pmatrix} XX^T & X\langle Z \rangle^T \\ \langle Z \rangle X^T & \langle ZZ^T \rangle \end{pmatrix} \right)^{-1},\\
\mu_{w_j^m} &= \langle (\Psi^m)_{j,:}^{-1} \rangle Y^m \begin{pmatrix} X^T & Z^T \end{pmatrix}\\
&\quad - \sum_{l \neq j} \langle (\Psi^m)_{j,l}^{-1} \rangle \langle W_{l,:}^m \rangle \begin{pmatrix} XX^T & X\langle Z \rangle^T \\ \langle Z \rangle X^T & \langle ZZ^T \rangle \end{pmatrix},\\
a_m &= a_0 + d_m / 2,\\
b_{mk} &= b_0 + \langle \| W_{:,k}^m \| \rangle / 2,
\end{aligned} \tag{17}$$

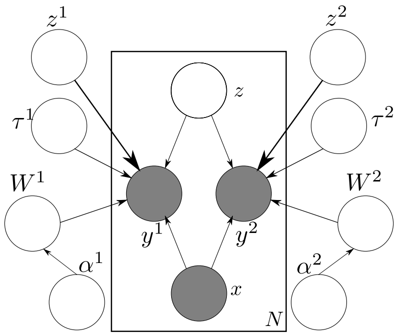


Figure 4. Graphical model for GSPCCA.

where $\mathrm{diag}\langle\alpha^m\rangle$ is the diagonal matrix with k-th element $\langle\alpha_k^m\rangle$.

4.2. Model with Isotropic Noise

The model proposed in the previous subsection requires a large number of calculations to infer noise precision matrices. Additionally, the prior distribution has a large influence when there are a small number of samples, because $\nu_0^m > d_m - 1$. Therefore, following the approach used by Klami et al. (2013), we propose a model that uses isotropic noise and non-shared latent variables. The generative model is

$$\begin{aligned} z_n &\sim \mathcal{N}(0, I_{d_z}), \\ z_n^m &\sim \mathcal{N}(0, I_{d_{z_m}}), \\ y_n^m &\sim \mathcal{N}\left(W_x^m x_n + A^m z_n + B^m z_n^m, (\tau^m)^{-1} I_{d_m}\right). \end{aligned} \tag{18}$$

When the z_n^m are integrated out, this model is equivalent to the model proposed in the previous subsection with $\Psi^m = B^m(B^m)^T + (\tau^m)^{-1} I_{d_m}$. So we can consider this model as equivalent to imposing a low-rank assumption on the covariance matrices. To simultaneously estimate A and B, we write $W_z = \begin{pmatrix} A^{(1)} & B^{(1)} & 0 \\ A^{(1)} & 0 & B^{(1)} \end{pmatrix}$, $W = \begin{pmatrix} W_x & W_z \end{pmatrix}$, and consider the model

$$\begin{aligned} \alpha_k^m &\sim \mathrm{Gamma}(a_0, b_0), \\ W_{:,k}^m &\sim \mathcal{N}(0, (\alpha_k^m)^{-1} I_{d_m}), \\ \tau^m &\sim \mathrm{Gamma}(a_0, b_0), \\ z_n &\sim \mathcal{N}(0, I_{(d_z + d_{z_1} + d_{z_2})}), \\ y_n^m &\sim \mathcal{N}\left(W_x^m x_n + W_z^m z_n, (\tau^m)^{-1} I_{d_m}\right), \end{aligned} \tag{19}$$

as shown in Figure 4. This representation reduces the number of model parameters. We refer to this model as group sparse PCCA (GSPCCA). This model also requires small hyperparameters. We have used $a_0, b_0 = 10^{-14}$ in our experiments. Additionally, we choose the approximate posterior

$$q(Z, \Theta) = q(Z) \prod_m \left(q\left(\tau^m\right) q\left(\alpha^m\right) q\left(W^m\right)\right), \tag{20}$$

and the shape

$$\begin{aligned} q(Z) &= \prod_n \mathcal{N}(\mu_{z_n}, \Sigma_z), \\ q(W^m) &= \prod_d \mathcal{N}(\mu_{W_{d,:}^m}, \Sigma_{W^m}), \\ q(\alpha^m) &= \prod_k \mathrm{Gamma}(a_{\alpha^m}, b_{\alpha_k^m}), \\ q(\tau^m) &= \mathrm{Gamma}(a_{\tau^m}, b_{\tau^m}). \end{aligned} \tag{21}$$

The parameters are updated as

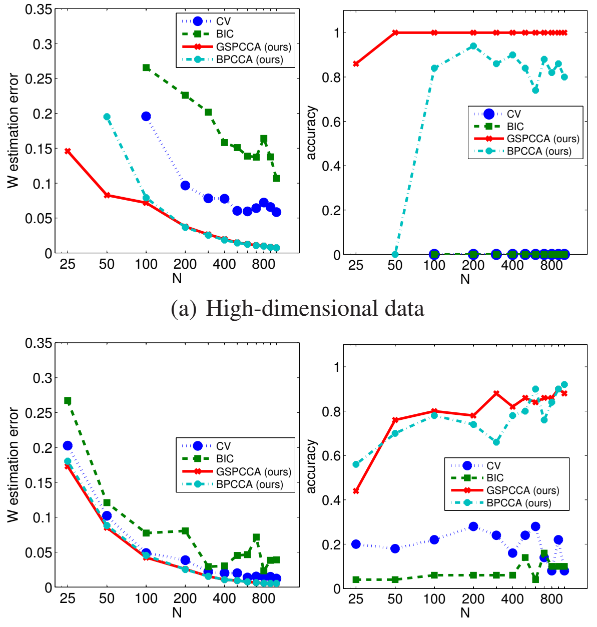
$$\begin{aligned} \Sigma_{W^m} &= \left(\mathrm{diag}\langle\alpha^m\rangle + \langle\tau^m\rangle \begin{pmatrix} XX^T & X\langle Z\rangle^T \\ \langle Z\rangle X^T & \langle ZZ^T\rangle \end{pmatrix}\right)^{-1}, \\ \mu_{W^m} &= Y^m \begin{pmatrix} X^T & \langle Z^T\rangle \end{pmatrix}, \\ \Sigma_z &= \left(I + \sum_m \langle\tau^m\rangle \langle (W_z^m)^T W_z^m\rangle\right)^{-1}, \\ \langle Z\rangle &= \Sigma_z \left(\sum_m \langle\tau^m\rangle \left(\langle (W_z^m)^T\rangle Y^m - \langle (W_z^m)^T W_x^m\rangle X\right)\right), \\ a_{\alpha^m} &= a_0 + d_m/2, \\ b_{\alpha_k^m} &= b_0 + \langle (W^m)^T W^m\rangle_{k,k}/2, \\ a_{\tau^m} &= a_0 + N d_m/2, \\ b_{\tau^m} &= b_0 + \frac{1}{2}\Bigg(\mathrm{Tr}\left(Y^m (Y^m)^T\right. \\ &\quad - 2Y^m \begin{pmatrix} X^T & \langle Z^T\rangle \end{pmatrix} \langle (W^m)^T\rangle\big) \\ &\quad + \mathrm{Tr}\left(\langle (W^m)^T W^m\rangle \begin{pmatrix} XX^T & X\langle Z\rangle^T \\ \langle Z\rangle X^T & \langle ZZ^T\rangle \end{pmatrix}\right)\Bigg). \end{aligned} \tag{22}$$

4.3. Optimization of the Linear Transformation of the Latent Variables

The maximum likelihood solution of probabilistic partial CCA has the same degrees of freedom as the linear transformation of latent variables. In the Bayesian model, we optimize this transformation in each iteration to obtain an approximate distribution that is closer to the prior distribution. We expect that this speeds up the convergence and that the latent variables are more independent. The function to be maximized is similar to that in (Virtanen et al., 2011), and is defined as

$$\begin{aligned} L(R) &= -\frac{\mathrm{Tr}(R^{-1}\langle ZZ^T\rangle R^{-T})}{2} + (d_1 + d_2 - N)\log|R| \\ &\quad - \frac{1}{2}\sum_{m=1}^{2} d_m \sum_{k=1}^{d_z} \log(r_k^T \langle (W_z^m)^T W_z^m\rangle r_k). \end{aligned} \tag{23}$$

To solve this, we use the L-BFGS method (Liu & Nocedal, 1989) initialized with the identity matrix. Using the opti-



(a) High-dimensional data

(b) Low-dimensional data

Figure 5. Comparison of the W_x estimation error and the model accuracy. The left panel shows the relative estimation error of W_x. The right panel shows the accuracy of d_z.

mal R, the approximate distributions are transformed into

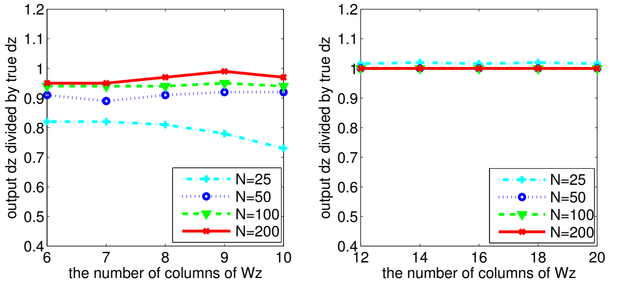
$$\begin{aligned} \langle Z \rangle &\leftarrow R^{-1}\langle Z \rangle, \\ \Sigma_Z &\leftarrow R^{-1}\Sigma_Z R^{-T}, \\ \mu_{W_z^m} &\leftarrow \mu_{W_z^m} R, \\ \Sigma_{W_z^m} &\leftarrow R^T \Sigma_{W_z^m} R. \end{aligned} \tag{24}$$

5. Experiments

We have applied our methods to synthetic and real data, to verify that they can be used with a small number of samples or high-dimensional data. We compared the stability of the model selection and the causality measures.

5.1. Model Selection

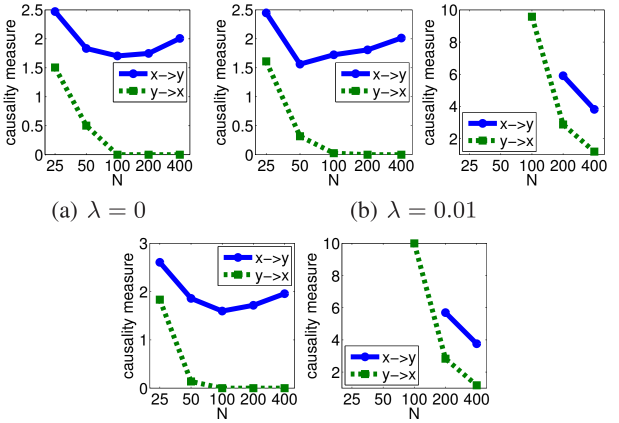
We first investigated the estimates of W_x and d_z using synthetic data. We did not consider W_z because the maximum likelihood solution of W_z is not unique. We compared our methods (BPCCA, GSPCCA) with the model selection techniques using the Bayesian information criterion (BIC) and five-fold cross validation (CV). In our methods, we considered that a component k of the solution was active when $\langle \alpha_k^m \rangle < 50$, and let d_z be an estimate of the number of k for that are active for each view. We set $d_1 = 5, d_2 = 4, d_x = 3$, and $d_z = 2$ for low-dimensional data, and $d_1 = 50, d_2 = 50, d_x = 5$, and $d_z = 5$ for high-dimensional data. In each setting, we generated $N = 25, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000$ samples from a generative model. Each column of the projection matrix was sampled from a Normal distribution with zero mean and unit variance, and the noise covariance matrices were $I_{d_z} + \sum_{i=1}^{\lfloor \frac{d_z}{2} \rfloor} u_i u_i^T$, for $u_i \sim \mathcal{N}(0, I_{d_z})$. W_z had five columns for low-dimensional data, and 10 columns for high-dimensional data. We conducted 50 experiments for each parameter. For the Bayesian methods, we determined that the method had converged if the relative change in the variational lower bound was below 10^{-4}. As it converges to a local maxima, we initialized the model by randomly sampling the latent variables from the prior, and ran the algorithm 10 times choosing the solution with the best variational lower bound. For each method, we calculated the mean of the relative error of W_x using



(a) low-dimensional data

(b) high-dimensional data

Figure 6. Comparison of the estimates of d_z. The left panel shows the performance on low-dimensional data. The right panel corresponds to high-dimensional data.



(a) $\lambda = 0$

(b) $\lambda = 0.01$

(c) $\lambda = 0.1$

Figure 7. Comparison of the stability of a causality measure. (a) and the left panels of (b) and (c) show the performance of GSPCCA (ours). The right panels of (b) and (c) show the performance of PCCA. The blue line shows the estimated causality measures for the true direction. The green line shows the estimates for the reverse direction.

$\frac{\mathrm{Tr}((W_x-\hat{W}_x)^T(W_x-\hat{W}_x))}{\mathrm{Tr}(W_x^T W_x)}$, where $\hat{W}_x$ is an estimate of W_x. We also recorded the accuracy rate of the system, d_z.

The results are presented in Figure 5. In the right subfigure of Figure 5(a), the CV result is hidden because it has been overwritten by the BIC result. Because we cannot stably calculate the BIC and CV of non-Bayesian methods when $D = 50$ and $N = 25, 50$, and the BIC and CV of BPCCA cannot be calculated when $D = 50$ and $N = 25$, we have not included these results. These plots show that existing model selection methods perform poorly and that the accuracy decreases to zero in high dimensions. Conversely, the two Bayesian methods are very accurate, even for high-dimensional data. BPCCA's performance degrades when $D = 50$ and $N = 50$, but GSPCCA's performance degrades more gradually. The estimate of W_x follows a similar trend. These results demonstrate that our methods calculate the model selection and parameter estimation more accurately than non-Bayesian methods, and that GSPCCA is the best method.

Next, we compared the model-selection performance of GSPCCA by varying the number of columns of W_z to 6, 7, 8, 9, and 10 for low-dimensional data, and 12, 14, 16, 18, 20 for high-dimensional data, with $N = 25, 50, 200, 800$. The performance was measured using the mean of the number of active components divided by the true d_z. The results are shown in Figure 6. In high dimensions, the performance is almost one for all the parameters. In low dimensions, if $N = 25$ the performance decreases gradually. However, this effect can be ignored because the true d_z is two. These results indicate that the number of columns in W_z has little effect on the performance, if it is sufficiently large.

5.2. Causality Measure with Synthetic Data

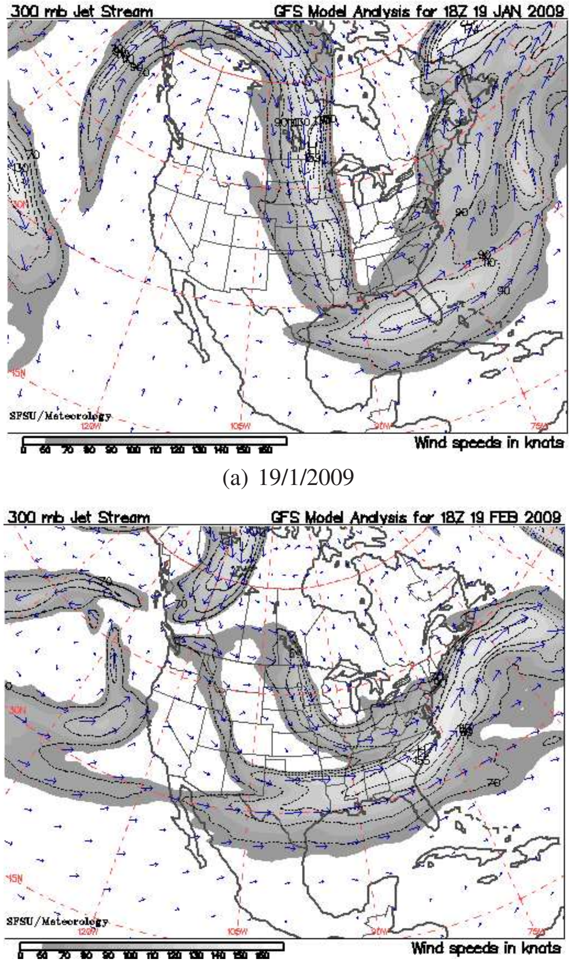
To evaluate the stability of the causality calculations for a small sample of high-dimensional data, we generated a time series using the following linear model.

$$\begin{aligned} x_t &= 0.5x_{t-1} + \epsilon_{t,x}, \\ y_{2_t} &= 0.5y_{2_{t-1}} + Wx_{t-1} + \epsilon_{t,y_2}, \\ y_t &= \begin{pmatrix} y_{2_t}^T & y_{2_t}^T \end{pmatrix}^T + \epsilon_{t,y}, \end{aligned} \tag{25}$$

where the first two columns of W are sampled from $\mathcal{N}(0, 0.5 \cdot I_{20})$ and the other columns are zero. $\epsilon_{t,x}, \epsilon_{t,y_2}$ denotes Gaussian noise with zero mean and unit variance. $\epsilon_{t,y}$ is 0 when $r = 0$, and is Gaussian noise with zero mean and variance $r \cdot I_{40}$ otherwise. The true causality direction is $x \rightarrow y$. The first and second halves of y_t are strongly correlated. This correlation is strong when r is small. The optimal dimension of the latent variables is two. Using this model, we set the embedding dimension to 1, r to $0, 0.01, 0.1$, and the sample size to $N = 25, 50, 100, 200, 400$, for each parameter. We expect that the causality measures derived from PCCA and probabilistic PCCA are equivalent, so we compared PCCA with GSPCCA (the best performing method). We used $\sum_{d=1}^{20} \frac{1}{2} \log_2 \frac{1}{1-\rho_d^2}$ as a causality measure for PCCA. For GSPCCA, we let ρ_k be the correlation between $\langle Y_{t-1(k,:)} | Y_t \rangle$ and $\langle Y_{t-1(k,:)} | X_{t-1} \rangle$, and used $\sum_k \frac{1}{2} \log_2 \frac{1}{1-\rho_k^2}$ as a causality measure, where the summation is over the active components. Figure 7 shows the results. We have not included results if the solution could not be stably evaluated. The causality measure using PCCA diverged when N was below 200, irrespective of λ. This measure also increased in the direction of $y \rightarrow x$, so this measure is unreliable when N is small. However, the measure using GSPCCA was zero in the $y \rightarrow x$ direction when N was larger than 100, because the Bayesian model makes directions that have a negligible influence converge to zero. This behavior helps eliminate false causality relations, but this model may overlook true causality relations when the influence is small. In such cases, we could detect small influences by modifying the hyperparameters of the ARD prior. This measure tended to diverge when λ was less than 0.01 and $N = 50$, or $\lambda = 0.1$ and $N = 25$. However, the measure in the $x \rightarrow y$ direction was larger than that in the $y \rightarrow x$ direction. The Bayes model also becomes unstable when there are an insufficient number of samples.

5.3. Causality Measure with Real Data

Next, we applied GSPCCA and PCCA to meteorological data, using the Global Summary of the Day (GSOD) provided by the National Climatic Data Center (NCDC) on its website. For this experiment, we used data from the USA between December 24, 2008 and February 28, 2009. Figure 8 shows the observed jet stream during that same period. We selected seven types of variables that did not have a substantial amount of missing data: mean temperature, mean dew point, mean visibility, mean wind speed, maximum sustained wind speed, maximum temperature, and minimum temperature. Therefore, the time series has seven dimensions. The length was 66. We randomly chose 224 targets based on distance, after excluding targets with many missing values. We conducted a zero-order hold for missing values. We set the embedding dimensions to 2, 3, and 4 and used the same causality measure as in the synthetic data experiments. Figure 9 shows our results. Figure 9 shows the largest 50 index values. Because the causality measure that used PCCA with the embedding dimension of four diverged in some pairs, we have included all the index values that diverged. When the embedding dimension was two, GSPCCA and PCCA had a similar tendency to show a strong information flow from west to east over the eastern region, and from north to south in the central region. This is consistent with Figure 8. When the embedding dimensions were four, the arrows drawn using PCCA



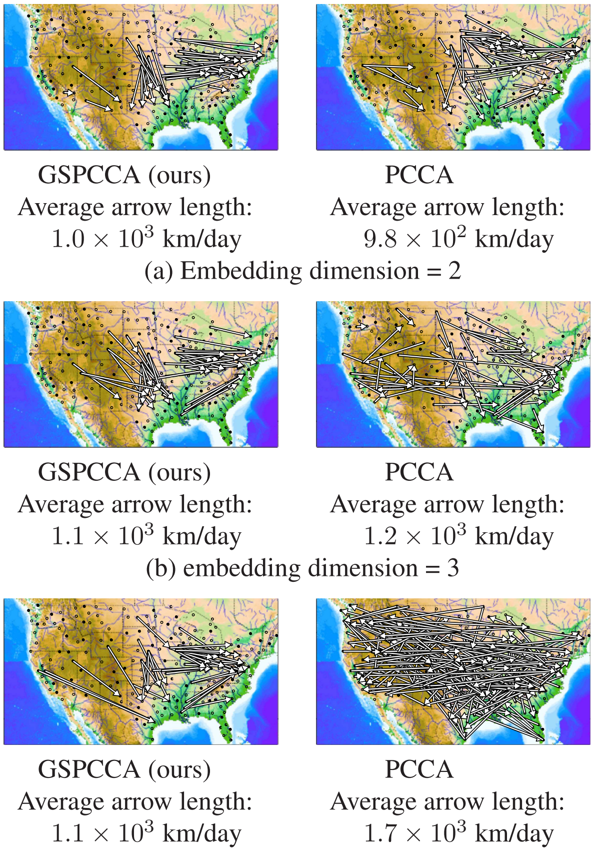
(a) 19/1/2009

(b) 19/2/2009

Figure 8. Weather information flow map for the USA (source: The California Regional Weather Server, San Francisco University).

were scattered over the mainland, although the index values using GSPCCA had a similar tendency to those with the embedding dimension of two. This result implies that PCCA overfits the data when the embedding dimension is high.

Next, we calculated the average arrow length using the Hubeny formula[1]. It was $1.0 \times 10^3, 1.1 \times 10^3, 1.1 \times 10^3$ km/day for GSPCCA and $9.8 \times 10^2, 1.2 \times 10^3, 1.7 \times 10^3$ km/day for PCCA. This shows that the causality measure using GSPCCA was more stable and similar to the actual air current, which was approximately 8.6×10^2 km/day (Shibuya et al., 2011), even when the embedding dimension was high. Because the true embedding dimension is unknown, GSPCCA is a more reliable method.



GSPCCA (ours)
Average arrow length:
1.0×10^3 km/day

PCCA
Average arrow length:
9.8×10^2 km/day

(a) Embedding dimension = 2

GSPCCA (ours)
Average arrow length:
1.1×10^3 km/day

PCCA
Average arrow length:
1.2×10^3 km/day

(b) embedding dimension = 3

GSPCCA (ours)
Average arrow length:
1.1×10^3 km/day

PCCA
Average arrow length:
1.7×10^3 km/day

(c) Embedding dimension = 4

Figure 9. Weather information flow map of USA (2008/12/24 – 2009/02/28). Maps on the left were calculated using GSPCCA, and maps on the right were calculated using PCCA.

6. Conclusion

We proposed a probabilistic interpretation of partial CCA. We also presented a Bayesian extension and an inference algorithm based on the probabilistic interpretation. Our experiments have demonstrated that the proposed methods are more appropriate for model selection and estimating causal relations from time series than existing methods, when there are a small number of samples or in high dimensions. We expect that PCCA and causality measures will be extensively applied to many areas using our methods.

Our Bayesian partial CCA method can be extended to a robust estimation method using a Student distribution for the noise (Archambeau et al., 2006), or to an inference method using the online variational Bayes technique (Hoffman et al., 2013). Additionally, by considering the projection matrices as random variables, we can construct a more complex model that allows the causal relation to change over time.

[1] `http://www.kashmir3d.com/kash/manual-e/std_siki.htm`

References

Archambeau, Cédric, Delannay, Nicolas, and Verleysen, Michel. Robust probabilistic projections. In *ICML*, pp. 33–40, 2006.

Bach, Francis R and Jordan, Michael I. A probabilistic interpretation of canonical correlation analysis. Technical Report 688, Department of Statistics, University of California, Berkeley, 2005.

Chávez, Mario, Martinerie, Jacques, and Le Van Quyen, Michel. Statistical assessment of nonlinear causality: application to epileptic eeg signals. *Journal of Neuroscience Methods*, 124(2):113–128, 2003.

Damianou, Andreas, Ek, Carl, Titsias, Michalis K, and Lawrence, Neil D. Manifold relevance determination. In *ICML*, pp. 145–152, 2012.

Ek, Carl Henrik, Rihan, Jon, Torr, Philip HS, Rogez, Grégory, and Lawrence, Neil D. Ambiguity modeling in latent spaces. In *MLMI*, pp. 62–73, 2008.

Fujiwara, Yusuke, Miyawaki, Yoichi, and Kamitani, Yukiyasu. Estimating image bases for visual image reconstruction from human brain activity. In *NIPS*, pp. 576–584, 2009.

Granger, Clive WJ. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 37(3):424–438, 1969.

Hardoon, David R and Shawe-Taylor, John. Sparse canonical correlation analysis. *stat*, 1050:19, 2009.

Hoffman, M, Blei, D, Wang, Chong, and Paisley, John. Stochastic variational inference. *JMLR*, 14:1303–1347, 2013.

Hotelling, Harold. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Klami, Arto and Kaski, Samuel. Local dependent components. In *ICML*, pp. 425–432, 2007.

Klami, Arto, Virtanen, Seppo, and Kaski, Samuel. Bayesian canonical correlation analysis. *JMLR*, 14:965–1003, 2013.

Kowalski, J, Tu, XM, Jia, G, Perlis, M, Frank, E, Crits-Christoph, P, and Kupfer, DJ. Generalized covariance-adjusted canonical correlation analysis with application to psychiatry. *Statistics in medicine*, 22(4):595–610, 2003.

Lai, Pei Ling and Fyfe, Colin. Kernel and nonlinear canonical correlation analysis. *IJNS*, 10(05):365–377, 2000.

Leen, Gayle and Fyfe, Colin. A gaussian process latent variable model formulation of canonical correlation analysis. In *ESANN*, pp. 413–418, 2006.

Liu, Dong C and Nocedal, Jorge. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

Melzer, Thomas, Reiter, Michael, and Bischof, Horst. Kernel canonical correlation analysis. In *ICANN*, pp. 353–360, 2001.

Neal, Radford M. *Bayesian learning for neural networks*. PhD thesis, University of Toronto, 1995.

Rao, B Raja. Partial canonical correlations. *Trabajos de estadistica y de investigación operativa*, 20(2):211–219, 1969.

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

Shibuya, Takashi, Harada, Tatsuya, and Kuniyoshi, Yasuo. Causality quantification and its applications: structuring and modeling of multivariate time series. In *KDD*, pp. 787–796, 2009.

Shibuya, Takashi, Harada, Tatsuya, and Kuniyoshi, Yasuo. Reliable index for measuring information flow. *Physical Review E*, 84(6):061109, 2011.

Sumioka, Hidenobu, Yoshikawa, Yuichiro, and Asada, Minoru. Development of joint attention related actions based on reproducing interaction contingency. In *ICDL*, pp. 256–261, 2008.

Verdes, PF. Assessing causality from multivariate time series. *Physical Review E*, 72(2):026222.1–026222.9, 2005.

Vía, Javier, Santamaría, Ignacio, and Pérez, Jesús. A learning algorithm for adaptive canonical correlation analysis of several data sets. *Neural Networks*, 20(1):139–152, 2007.

Virtanen, Seppo, Klami, Arto, and Kaski, Samuel. Bayesian cca via group sparsity. In *ICML*, pp. 457–464, 2011.

Wang, Chong. Variational bayesian approach to canonical correlation analysis. *IEEE Transactions on Neural Networks*, 18(3):905–910, 2007.

Yamashita, Yuya, Harada, Tatsuya, and Kuniyoshi, Yasuo. Causal flow. *IEEE Transactions on Multimedia*, 3(3): 619–629, 2012.

Yger, Florian, Berar, Maxime, Gasso, Gilles, and Rakotomamonjy, Alain. Adaptive canonical correlation analysis based on matrix manifolds. In *ICML*, pp. 1071–1078, 2012.