Continuum Power CCA: A Unified Approach for Isolating Coupled Modes

ERIK SWENSON*
Asia–Pacific Economic Cooperation Climate Center, Busan, South Korea

(Manuscript received 29 June 2014, in final form 21 October 2014)

ABSTRACT

Various multivariate statistical methods exist for analyzing covariance and isolating linear relationships between datasets. The most popular linear methods are based on singular value decomposition (SVD) and include canonical correlation analysis (CCA), maximum covariance analysis (MCA), and redundancy analysis (RDA). In this study, continuum power CCA (CPCCA) is introduced as one extension of continuum power regression for isolating pairs of coupled patterns whose temporal variation maximizes the squared covariance between partially whitened variables. Similar to the whitening transformation, the partial whitening transformation acts to decorrelate individual variables but only to a partial degree with the added benefit of preconditioning sample covariance matrices prior to inversion, providing a more accurate estimate of the population covariance. CPCCA is a unified approach in the sense that the full range of solutions bridges CCA, MCA, RDA, and principal component regression (PCR). Recommended CPCCA solutions include a regularization for CCA, a variance bias correction for MCA, and a regularization for RDA. Applied to synthetic data samples, such solutions yield relatively higher skill in isolating known coupled modes embedded in noise. Provided with some crude prior expectation of the signal-to-noise ratio, the use of asymmetric CPCCA solutions may be justifiable and beneficial. An objective parameter choice is offered for regularization with CPCCA based on the covariance estimate of O. Ledoit and M. Wolf, and the results are quite robust. CPCCA is encouraged for a range of applications.

1. Introduction

Various multivariate statistical methods have been developed and proven useful for analyzing covariance and estimating physically meaningful linear relationships between multiple coupled variables. Such approaches have many applications spanning a broad range of disciplines, although they are particularly useful for diagnosis and prediction of the physical climate system. Linear methods tend to use singular value decomposition (SVD) to identify a related subspace with projection vectors that maximize some measure of relation. Two popular and well-understood symmetric methods are canonical correlation analysis (CCA; Hotelling 1936) and maximum covariance analysis (MCA; Wallace et al. 1992), which generate pairs of stationary spatial patterns whose temporal covariation produces the highest possible squared correlation and squared covariance, respectively. In particular, CCA has been used to diagnose and predict land air temperature based on covarying sea surface temperature (SST) patterns (Barnett and Preisendorfer 1987; DelSole and Shukla 2006; Barnston and Smith 1996). Similarly, MCA has been used to infer teleconnections with SST (Wallace et al. 1992; Czaja and Frankignoul 2002; Frankignoul and Sennéchael 2007; Wu 2010) and for statistical downscaling (Tippett et al. 2008; Tung et al. 2013). For diagnosing covariability, Navarra and Tribbia (2005) establish a more unified framework. Alternatively, asymmetric approaches may be useful and include redundancy analysis (RDA; von Storch and Zwiets 1999) and partial least squares (PLS) regression (Wold 1966), which maximize the explained variance of one variable by another. These approaches have also been applied for diagnosis (e.g., Bakalian et al. 2010) and prediction (e.g., Smoliak et al. 2010).

Following a comparison study by Bretherton et al. (1992), there has been a drive to improve theoretical understanding regarding the aspects and limitations of CCA and MCA. Both approaches (and many others) have a common basis in linear regression (Tippett et al. 2008). Drawbacks of MCA are its representation of relationships...
with transformations that are necessarily orthogonal (Newman and Sardeshmukh 1995), its lack of a least squares estimate because of nonzero correlation between modes (Tippett et al. 2008), and its bias toward modes with larger variance given that it maximizes squared covariance. Still, this bias can lead to more stable solutions, and a strong appeal of MCA is in how it avoids inverting sample covariance matrices, something necessary for CCA. This can lead to dramatic overfitting, particularly when applied to small sample sizes that have high colinearity (among predictors or spatial points), properties characteristic of typical climatic data. This is because noise (unrelated variation in the data) can be dramatically inflated and strongly affect the leading modes if it by chance increases the maximized quantity. Consequently, CCA and MCA will identify spurious relationships when applied to noise alone (Cherry 1996), and verification in independent data is generally the fairest indication of overfitting. Despite often being plagued by overfitting in small samples, CCA is quite theoretically attractive in that it generates a least squares estimate with uncorrelated modes, as well as providing an optimal noise filter for autoregressive models (DelSole and Chang 2003). In practice, CCA and related techniques simply require an additional filter, as discussed next.

In efforts to filter out noise by preconditioning sample covariance matrices, regularization or predictor reduction techniques have been applied to existing methods at the expense of losing optimality across the sample. For CCA, approaches include regularized CCA (RCCA; Vinod 1976) and penalized CCA (Hastie et al. 1995), which both use ridge regression to apply spatial smoothing, as well as kernel CCA (Akaho 2001) and CCA using truncation according to empirical orthogonal functions (EOFs) [Barnett–Preisendorfer (BP) CCA; Barnett and Preisendorfer 1987]. Unfortunately, solutions of BP CCA are often highly sensitive to the number of EOFs retained, and potentially important relationship information is entirely neglected if present in trailing EOFs. Given that the leading (trailing) EOFs of autocorrelated variables tend to span large (small) spatial scales, RCCA can also tend to damp higher-order EOFs yet in a much smoother fashion. In the face of overfitting, the primary concern with applying RCCA and BP CCA is establishing an objective basis for choosing the ridge parameter or truncation number, respectively. This is an issue for any regularization method. Cross validation is effective but may not be possible with small samples given that sample data often cannot be used for both choosing the parameter and applying the approach. For BP CCA, measures such as Akaike’s information criterion may be computed. Following the derivation of an objective and optimal choice for the ridge parameter by Ledoit and Wolf (2004) with respect to covariance estimation, application to RCCA has yielded robust results (Cruz-Cano and Lee 2014).

Regularization techniques are attractive because they can handle missing data and often yield more robust results than conventional techniques, particularly when the number of predictors (spatial points) exceeds the sample size. This is common in climatic data, and multiple climate studies have benefited from using regularization. To name a few, for paleoclimate reconstruction, Smerdon et al. (2010) evaluate BP CCA alongside regularized expectation maximization (Schneider 2001), and Tingley et al. (2012) summarize various approaches and the important issues involved. For climate change detection, Ribes et al. (2009) apply ridge regression to the optimal fingerprint method using the Ledoit–Wolf covariance estimate (Ledoit and Wolf 2004). More recently, for climate prediction, Lim et al. (2011) improve seasonal rainfall prediction using RCCA. Also, Fischer (2014) regularizes principal covariates regression (de Jong and Kiers 1992), yielding a continuum approach that links principal component regression (PCR) with RDA and PLS regression, and further applies it to identify coupled patterns between seasonal rainfall and sea level pressure.

An alternative regularization technique that has yet to be applied in this context is continuum power regression (Lorber et al. 1987; Stone and Brooks 1990), an approach that bridges ordinary least squares, PLS regression, and PCR. Partial whitening, the transformation involved with continuum power regression, acts to decorrelate differing spatial points, but only to a partial degree specified by a parameter. Here it is applied as a regularization for sample covariance matrices prior to computing CCA, yielding a new approach for linearly relating variables that may be referred to as continuum power CCA (CPCCA). CPCCA maximizes the squared cross covariance between partially whitened variables and provides a unique plane of solutions that bridges those of CCA, MCA, and RDA, as well as PCR in its upper limit. In terms of partial whitening parameters $\alpha$ and $\beta$, recommended solutions are a regularization for CCA (typically small $\alpha$ and $\beta$), a variance bias correction for MCA ($\alpha = \beta = 0.5$), and a regularization for RDA (typically small $\alpha$ with $\beta = 1$). An objective choice for regularization is proposed based on Ledoit and Wolf (2004). These solutions are demonstrated to skillfully isolate a coupled mode. Further, it is found that asymmetric solutions ($\alpha \neq \beta$) are justifiable based on prior expectation of asymmetry in the signal-to-noise ratio.

This paper is organized as follows: The partial whitening transformation is discussed in section 2. Section 3 presents the formulation of CPCCA and the associated basis set properties. In section 4, CPCCA is
applied to synthetic data for isolating coupled modes embedded in noise. Results are contrasted with those of CCA, MCA, RDA, and RCCA, and an objective parameter choice is proposed for regularization. Finally, section 5 provides a summary and conclusions.

2. The partial whitening transformation

A common and useful operation applied to a univariate normally distributed stationary process is normalization, which results from dividing by the square root of the sample variance or standard deviation. This operation scales a process so that it has unit variance. Now suppose instead that one divides by the sample standard deviation taken to a positive fractional power between 0 and 1. Such a partial normalization may be alternatively used to scale the process such that its sample variance lies between 1 and the original variance.

Partial normalization may be generalized to multiple dimensions in terms of matrix operations, but first it is necessary to discuss covariance and the whitening transformation. In a multivariate framework let the term “covariance” refer to the sample covariance of a variable with itself, and let the term “cross covariance” refer to the sample covariance between two different variables. Provided with variables \( \mathbf{X} \) and \( \mathbf{Y} \), matrices of anomalies (zero sample mean for each spatial point) with rows giving different spatial points and columns giving different samples (e.g., temporal variation), assume that spatial points have equal weighting (e.g., uniformly spaced grids) and have absorbed a factor of \((n - 1)^{-1/2}\), with \(n\) giving the number of samples. Then, individual covariances for \( \mathbf{X} \) and \( \mathbf{Y} \) are given simply by \( \mathbf{XX}^T \) and \( \mathbf{YY}^T \), respectively, and cross covariance is given by \( \mathbf{XY}^T \). For \( \mathbf{X} \), values of diagonal (off diagonal) elements of the symmetric positive definite matrix \( \mathbf{XX}^T \) represent the sample variances at individual spatial points (covariances between differing spatial points). The corresponding standard deviation matrix results from applying a matrix square root, yielding \((\mathbf{XX}^T)^{1/2}\), defined by \( \mathbf{XX}^T = (\mathbf{XX}^T)^{1/2}(\mathbf{XX}^T)^{1/2} \).

The whitening transformation (DelSole and Tippett 2007) is a particularly useful transformation that involves multiplying a variable with the inverse of the matrix square root of its covariance matrix, for \( \mathbf{X} \) given by \((\mathbf{XX}^T)^{-1/2}\mathbf{X} \). This operation both decorrelates differing spatial points and normalizes individual spatial points, given that the sample covariance matrix of a whitened variable is equivalent to the identity matrix, with \((\mathbf{XX}^T)^{-1/2}\mathbf{XX}^T(\mathbf{XX}^T)^{-1/2} = \mathbf{I} \). Note that, when the number of spatial points exceeds that of samples \(n\), the inverse does not exist, although it may be replaced with the Moore–Penrose pseudoinverse. Whitening represents the multivariate form of normalization and is important in the context of regression in how it removes collinearity.

In practical applications with sample data, the full decorrelation with the whitening transformation typically comes at the expense of inflating noise when taking the inverse. For an ill-conditioned near-singular sample covariance matrix, such unidentifiable yet presumably small noise can even dominate the inverse. Regularization approaches have been used to alleviate noise inflation by preconditioning the sample covariance matrix at the expense of not completely removing collinearity. Such approaches include ridge regression (Vinod 1976; Hastie et al. 1995) and the use of EOF truncation (Barnett and Preisendorfer 1987). Another approach is continuum power regression (Lorber et al. 1987; Stone and Brooks 1990), which involves an operation that may be referred to as partial whitening, a transformation that only partially decorrelates and normalizes the data to a specified degree. For \( \mathbf{X} \) and \( \mathbf{Y} \), partial whitening is represented by transformation matrices \( \mathbf{T}_x \) and \( \mathbf{T}_y \), expressed as functions of parameters \( \alpha \) and \( \beta \), respectively, by

\[
\mathbf{T}_x = (\mathbf{XX}^T)^{(\alpha - 1)/2} \quad \text{and} \quad \mathbf{T}_y = (\mathbf{YY}^T)^{(\beta - 1)/2}. \tag{1}
\]

Note that parameter notation differs from that of Stone and Brooks (1990), and parameters may be reexpressed as any positive function. Applying partial whitening transformations in Eq. (1) to \( \mathbf{X} \) and \( \mathbf{Y} \) yields partially whitened variables \( \mathbf{X}^* = \mathbf{T}_x \mathbf{X} \) and \( \mathbf{Y}^* = \mathbf{T}_y \mathbf{Y} \), respectively. Generally, the term “partial whitening” refers to the use of fixed positive fractional parameter values \((0 < \alpha < 1 \text{ and } 0 < \beta < 1)\), although any positive values are valid in its form (including 0). For \( \mathbf{X} \), as \( \alpha \) increases from 0 to 1, \( \mathbf{T}_x \) continuously transitions from \((\mathbf{XX}^T)^{-1/2}\) (whitening) to the identity matrix (no transformation), respectively, given that \((\mathbf{XX}^T)^{0} = \mathbf{I}\). The subsequent effect of partial whitening transitions from the removal of all of the original collinearity to no change. Note that for symmetric positive definite matrices, a fractional power only deflates eigenvalues (while retaining eigenvectors). Given that the eigenvalues of \( \mathbf{XX}^T \) are the variances of the individual EOFs of \( \mathbf{X} \), partial whitening preserves EOFs while assigning exponents to EOF standard deviations, such that varying \( \alpha \) generally acts to vary the disparity in EOF weights (see appendix A). For \( \mathbf{X} \), as \( \alpha \) increases from 0 to 1, the weighting of the EOFs changes from being all equal to 1 (whitened) to being weighted by the original standard deviations (unchanged). Note that ridge regression (or Tikhonov regularization) has been similarly applied to the whitening transformation (e.g., for CCA; Vinod 1976) yet spans a linear path, distinct from that spanned by partial whitening (see section 4).
For $X$, whitening involves the inversion of $(XX^T)^{1/2}$, whereas partial whitening involves the inversion of $(XX^T)^{(1-\alpha)/2}$. Suppose that $(XX^T)^{1/2}$ has a condition number of $\kappa$, a value that determines the upper bound on noise inflation during inversion. It can then be easily shown that $(XX^T)^{(1-\alpha)/2}$ has a condition number of $\kappa^{1-\alpha}/\kappa$, which is always lower for $0 < \alpha < 2$, assuming that $\kappa > 1$.

From another perspective, partial whitening provides a means of shrinking the sample covariance matrix. Although the expectation of the sample covariance $XX^T$ is the population covariance, the population covariance matrix is often estimated more closely by $(XX^T)^{1-\alpha}$ (depending on the value of $\alpha$).

Partial whitening is distinguished here from a transformation under same terminology used by Donohue et al. (2007), which partially normalize harmonic amplitudes prior to analyzing coherence between multiple sound signals. Also using exponential scaling, partial whitening has a similar form yet acts to partially normalize EOFs. Also note that partial whitening is different from pseudowhitening, which only applies whitening to sound signals. Also using exponential scaling, partial whitening associated with fractional parameter $\alpha$.

Partial whitening provides a multivariate framework. The cross covariance between partially whitened variables may also be shown to represent a regression coefficient matrix involving transformed variables (see appendix B). This quantity is clarified further in the next section.

3. Continuum power CCA

In this section, CPCCA is introduced as a new unified or continuum approach for linearly relating two variables by isolating sets of multivariate relationships. Alongside CCA, MCA, and RDA, CPCCA is first derived as the solution to a constrained optimization problem. Solutions are then found in a more straightforward manner after the problem is expressed as MCA involving transformed variables. Following this, CPCCA basis sets are constructed and orthogonality conditions are examined. Finally, solutions are recommended in terms of $\alpha$ and $\beta$. Given computational constraints, CPCCA and related methods are ideally computed in principal component (PC) space for which CPCCA is derived in appendix A.

The traditional methods of CCA, MCA, and RDA may all be derived as optimization problems involving cross covariance $XY^T$ yet differing in terms of the set of norms that must be satisfied. CPCCA may be similarly formulated and represents a particular generalization. These approaches may be expressed by the optimization problem

$$\max_{q_x, q_y} (q_x^T XY^T q_y)$$

subject to

$$q_x^T XX^T q_x = 1 \quad \text{and} \quad q_y^T YY^T q_y = 1 \quad \text{for CCA-1},$$

$$q_x^T q_x = 1 \quad \text{and} \quad q_y^T q_y = 1 \quad \text{for MCA-1},$$

$$q_x^T XX^T q_x = 1 \quad \text{and} \quad q_y^T q_y = 1 \quad \text{for RDA-1},$$

or

$$q_x^T (XX^T)^{1-\alpha} q_x = 1 \quad \text{and} \quad q_y^T (YY^T)^{1-\beta} q_y = 1 \quad \text{for CPCCA-1}. (6)$$

The leading modes of CCA, MCA, and RDA result from finding projection vectors ($q_x$ and $q_y$) that are bounded by the different norms of Eqs. (3)–(5). Note that, only for MCA-1, projection vectors must have unit variance, whereas the temporally varying projections themselves ($X^T q_x$ and $Y^T q_y$) must have unit variance for CCA-1. More details regarding properties and solutions of CCA and MCA are discussed by Bretherton et al. (1992). CCA-1, MCA-1, and RDA-1 norms may also be recovered for different values of $\alpha$ and $\beta$ reflected by the norms for CPCCA-1 stated in Eq. (6). These norms may also be expressed in terms of the partial whitening transformation of Eq. (1), with $\tilde{q}_x^T \tilde{T}_s^{-2} q_x = 1$ and $\tilde{q}_y^T \tilde{T}_s^{-2} q_y = 1$. Solutions are most clearly found by reexpressing the problem in terms of $q_x^* = \tilde{T}_s^{-1} \tilde{q}_x$ and $q_y^* = \tilde{T}_s^{-1} \tilde{q}_y$, respectively. CPCCA-1 may then be formulated from MCA-1 involving partially whitened variables, the solution to the optimization problem

$$\max_{q_x^*, q_y^*} (q_x^*^T X^* Y^* T_{xy} q_y^*)$$

subject to

$$q_x^*^T q_x^* = 1 \quad \text{and} \quad q_y^*^T q_y^* = 1. \quad (8)$$

In the above form, the transformed solutions $q_x^*$ and $q_y^*$ must maximize the squared cross covariance between partially whitened variables while being bounded to have...
Solutions are found by decomposing $X^tY^st^T$ using SVD in terms of orthonormal singular vectors (columns of $U$ and $V$) and positive, sorted singular values (diagonal elements of $S$), with
\begin{equation}
X^tY^st^T = T_x^x T_y^y U^SV^T.
\end{equation}
Eqs. (7)–(11) may then be used to describe a broader class of approaches that involve transformation, computing SVD, followed by inverse transformation. For instance, RCCA solutions may be found by alternatively applying a ridge regression transformation (see section 4). The relationship with MCA is consistent the fact that CCA and RDA involve computing MCA with pre-whitened variables (Tippett et al. 2008). The critical distinction between different methods is the type of transformations that are used, which practically should constrain the solutions in some meaningful way. CPCCA offers a particular form based on covariance and justified as a regularization approach.

The traditional methods are special cases of CPCCA that satisfy orthogonality conditions. For $X$, this is demonstrated in that
\begin{equation}
P_x^TP_x = U^T (XX^T)^{1-\alpha} U = I \quad \text{when } \alpha = 1 \quad \text{(12)}
\end{equation}
\begin{equation}
R_y^TR_y = U^T (XX^T)^{\alpha} U = I \quad \text{when } \alpha = 0. \quad \text{(13)}
\end{equation}
Equation (12) states that orthonormal pattern vectors are preserved when no transformation is made, whereas Eq. (13) states that variates are restricted to be uncorrelated when whitening is performed to remove spatial autocorrelation.

For the weighted variates of $X$, when $\alpha > 0$ normalization is typically desired so that patterns absorb variance and may be represented with the same units as $X$. This is done through the normalization matrix $S_x^R = \text{Diag}(R_x^T R_x)^{1/2}$, where $\text{Diag} B$ denotes a diagonal matrix that contains the diagonal elements of $B$. Then the weighted patterns and normalized variates are given by the columns of $P_x^T S_x^R R_x^T$ and $R_y^TR_y S_y^R$, respectively, and similarly for $Y$. The cross-correlation matrix may then be represented by $C_{xy} = S_{x}^{R-1} (R_x^T R_y) S_{y}^{R-1}$, whereas the correlation between modes of $X$ may be represented by $C_{xx} = S_x^{R-1} (R_x^T R_x) S_x^{R-1}$ and similarly for $C_{yy}$. Examining different CPCCA modes $j$ and $k$ ($j \neq k$), given that differing modes have zero cross correlation ($C_{jk}^{xy} = 0$), it may be easily shown that differing modes for $X$ must satisfy $C_{jk}^{xy} \leq 1 - C_{jk}^{2}$, where $C_{jk}^{2}$ is the $i$th row and $j$th column of $C_{xy}$. Assuming a high but common value of $C_{jk}^{xy} = 0.95$, sample correlations between differing modes are restricted to be relatively small with $|C_{jk}^{xy}| \leq 0.3$. CPCCA variates of differing modes are nearly uncorrelated for very high cross-correlation values. Consistent with Eq. (13), this is typical for small $\alpha$, particularly when CPCCA is applied to small samples.

As is true for related approaches, keep in mind that orthogonality conditions affect modes trailing CPCCA-1, and this effect requires consideration if interpretation is
given to multiple CPCCA modes. Also note that, if
strictly uncorrelated variates are desired, the Gram–
Schmidt process (e.g., Navarra and Simoncini 2010) may
be applied to successively remove correlated portions of
the trailing CPCCA modes; however, this greatly in-
creases the computational expense.

Partial whitening parameters $\alpha$ and $\beta$ may be varied in
a continuous manner to produce a plane of CPCCA sol-
lutions that consecutively maximize the sample squared
cross covariance between partially whitened variables.
This quantity is essentially the multivariate form of the
univariate measure

$$\text{VAR}_x^a \times \text{COR}^2 \times \text{VAR}_y^\beta,$$  \hspace{1cm} (14)

where $\text{VAR}_x$ and $\text{VAR}_y$ are variances corresponding to $X$
and $Y$, respectively, and $\text{COR}$ is cross correlation. As
previously discussed, CPCCA bridges CCA ($\alpha = \beta = 0$),
MCA ($\alpha = \beta = 1$), and RDA ($\alpha = 0$ and $\beta = 1$ or vice
versa). Additionally, apart from special circumstances
(zero cross correlation between EOFs), CPCCA con-
verges to PCR in its upper limit yielding the EOFs of
individual variables (see appendix A). One way this may
be seen is that Eq. (14) is completely dominated by
$\text{VAR}_x^a$ or $\text{VAR}_y^\beta$ as either $\alpha$ or $\beta$ gets large, respectively.
Given the exponential form, CPCCA solutions tend to
resemble PCR for parameter values not much greater
than one, particularly if the value of the other parameter
is much less than 1. In this study, focus is given to solu-
tions that consecutively maximize the sample squared
cross covariance between partially whitened variables.

4. Synthetic experiments

In this section, CPCCA is examined in terms of the
expected skill in isolating a synthetically generated
coupled mode alongside the strength of the estimated
relationship as a function of partial whitening param-
eter values $\alpha$ and $\beta$. Experiments generally validate the
recommended solutions and justify the use of asym-
metry for different signal-to-noise ratios. Results are
generally comparable to RCCA. Contrasting CPCCA
solutions in an idealized framework provides some
objective support for the use of the particular recom-
ended solutions for relating more complex observed
linear relationships.

Synthetic datasets $X$ and $Y$, both with spatial dimension
$m = 60$ and number of time samples $n = 30$ (300 for some
cases), are constructed from a common set of pop-
ulation EOF patterns represented by spherical har-
monics. Specifically, population EOF-1 and EOF-2
patterns are sine and cosine waves, respectively, with half
of a period spanning the domain. Subsequent pairs of
EOFs of increasing wavenumbers ($\frac{1}{2}, \frac{5}{2}$, etc.) compose
population EOF-3 through EOF-10 and, by design, var-
iance is evenly distributed across space. The associated
variates follow random white noise with the expectation
to be uncorrelated, yet one true coupled mode is present
with a correlation of 0.9. This mode will be referred to as
the signal. The primary constraint for the signal is the
fractional weighting of population EOF-1 that is specified
differently for different experiments as $a$ and $b$ for $X$
and $Y$, respectively. With a total expected variance of 1,
population EOF-1 has a variance of 0.25 while the 10
leading EOFs together have a total variance of 0.8, with
individual variances decaying exponentially. In terms of
$\alpha$, the signal variance is given by $0.25a + 0.55(1-a)/9$ and
similarly for $b$. The population EOFs are added to ran-
domly generated background noise that follows red noise
in space [similar to Bretherton et al. (1992), using a spatial
autocorrelation of 0.9] and independent white noise in
time with a noise variance of 0.2. Further details for the
generation of synthetic datasets, including the process of
generating random data with an expected cross correla-
tion, are provided in appendix C.

Note that the idealized synthetic data are intended to
mimic the properties of interannually varying planetary-
scale geophysical variables of the observed climate sys-
tem (e.g., covariability between sea surface temperature
and 500-hPa geopotential height). The primary simplifications to consider when relating these results to real applications (e.g., the physical climate system) are linearity, the idealized spatial structure of the signal, and the presence of only one true coupled signal. Further discussion is made in section 5 regarding the caveats of applying CPCCA to real data and applicability of the synthetic results.

For signals of different magnitudes specified by \(a\) and \(b\), CPCCA-1 is computed with 100 synthetic realizations across a range of partial whitening parameters values varying evenly in increments of 0.05 for \(0 \leq a \leq 1.5\) and \(0 \leq b \leq 1.5\) spanning the solutions of CCA, MCA, and RDA. By design, these traditional methods maximize COR, squared covariance fraction (SCF), and the fraction of variance of \(Y\) explained by \(X\) [i.e., fraction of variance explained (FVE)], respectively, and these standard measures of relation are computed in order to assess the strength of the leading relationship represented by CPCCA-1. With CPCCA-1 represented in both space and time by \(X_{1}^{5p,1T}, X_{1}^{rT}, Y_{1}^{5p,1T}, Y_{1}^{rT}\), COR is computed by simply correlating the individual variates \(r_{x,1}\) and \(r_{y,1}\), whereas SCF and FVE are computed as a weighted mean-square error given by

\[
\text{SCF} = 1 - \| (X - X_{1}) (Y - Y_{1})^{T} \|_{F}^{2}/\|XX^{T}\|_{F}^{2} \quad \text{and} \quad \text{(15)}
\]

\[
\text{FVE} = 1 - \| (XX^{T})^{-1/2} (X - X_{1}) (Y - Y_{1})^{T} \|_{F}^{2}/\|XX^{T}\|_{F}^{-1/2}XY^{T}\|^{2} \quad \text{and} \quad \text{(16)}
\]

Note that, for any matrix \(B\), the square of the Frobenius norm, given by \(\|B\|_{F}^{2}\), is equivalent to the sum of the diagonal elements (or trace) of \(BB^{T}\). In addition to COR, SCF, and FVE, primary focus is given to the skill in isolating the true signal, given by \(X_{\text{signal}}\) and \(Y_{\text{signal}}\), which is computed similarly as the fraction of signal variance explained (FSVE) for individual variables \(X\) and \(Y\) (FSVE\(_{x}\) and FSVE\(_{y}\), respectively) given by

\[
\text{FSVE}_{x} = 1 - \|X_{\text{signal}} - X_{1}\|_{F}^{2}/\|X_{\text{signal}}\|_{F}^{2} \quad \text{and} \quad \text{(17)}
\]

\[
\text{FSVE}_{y} = 1 - \|Y_{\text{signal}} - Y_{1}\|_{F}^{2}/\|Y_{\text{signal}}\|_{F}^{2} \quad \text{and} \quad \text{(18)}
\]

As a whole, the skill in isolating the signal is then given simply by the average, with FSVE = (FSVE\(_{x}\) + FSVE\(_{y}\))/2.

Overfitting is accounted for by computing leave-one-out cross validation for FSVE and all measures of relation.

By definition, in the training data CCA-1, MCA-1, and RDA-1 always maximize COR, SCF, and FVE, respectively. However, the signal is more closely represented by CPCCA-1 using nonzero partial whitening parameter values generally \(\leq 0.5\). This is illustrated for three different signal magnitudes \((a = b = 1, a = b = 0.75, \text{and } a = b = 0.6)\) in Fig. 1, which shows the average FSVE (contours) computed for CPCCA-1 as a function of \(a\) and \(b\) alongside the coordinates of the maximum average FSVE, COR, SCF, and FVE. CPCCA-1 solutions using parameter values around 0.15–0.2 tend to be the most skillful. As the signal gets smaller, skill drops for all solutions. While MCA-1 performs relatively well

![Fig. 1. FSVE (contours at intervals of 0.05) for CPCCA-1 plotted as a function of partial whitening parameters \(\alpha\) and \(\beta\) averaged across 100 synthetic realizations for which the fractional weighting of EOF-1 for the signal in X and Y (\(a\) and \(b\), respectively) is specified as (a) 1, (b) 0.75, and (c) 0.6. Plotted are the \(\alpha\) vs \(\beta\) coordinates for maximum FSVE, COR, SCF, and FVE as indicated by an asterisk, solid dot, solid triangle, and solid square, respectively. The corresponding values are given above the panels.](image-url)
when the signal is population EOF-1 \((a = b = 1)\), its skill disappears when the signal gets smaller, whereas CPCCA-1 using small nonzero parameters remains moderately skillful. Solutions clearly diverge from the coupled signal when \(a\) or \(b\) is equal to 0, an expectation of CCA and RDA solutions applied to undersampled data without any type of regularization. Note that, after a truncation to around 10 EOFs, CCA-1 and RDA-1 solutions have much more comparable skill with those of CPCCA-1; however, some degree of partial whitening with CPCCA still yields relatively higher skill (not shown).

After cross validation, CPCCA-1 solutions using fractional parameter values remain with the highest relative skill (Fig. 2). The significant improvement over CCA-1 highlights CPCCA as a regularization approach. Also, the variance bias correction for MCA \((a = b = 0.5)\) is consistently a skillful choice. Within the range of solutions, some sensitivity to the true signal-to-noise ratio is apparent. As \(a\) and \(b\) are decreased from 1 to 0.75, the location of maximum FSVE (COR) shifts from \(a = b = 0.7\) (0.9) to about \(a = b = 0.35\) (0.4). This sensitivity is more apparent when there is an asymmetry in the signal strength (shown later). Note that solutions along the \(b = 1\) axis with nonzero \(a\) isolate the signal more accurately than RDA-1, and for smaller signals such solutions are nearly the most skillful for isolating the signal in \(Y\) alone (not shown). This reinforces CPCCA as a regularization for RDA. With \(a = b\), the signal itself is symmetric and therefore symmetric solutions tend to produce the highest values of FSVE, COR, and SCF. Note that CPCCA-1 is also computed for much larger sample sizes (not shown) and, as expected, skill is appreciably higher for all approaches. Differences between solutions are reduced; however, CPCCA-1 solutions using fractional parameter values still yield higher skill than CCA-1 and MCA-1.

In many practical situations, it is common for the signal-to-noise ratio of \(X\) to be appreciably different than that for \(Y\), considering that the nature of background noise may be fundamentally different. To address this, sets of synthetic experiments are repeated for \(a \neq b\): specifically, for \(a = 1, b = 0.75; a = 0.75, b = 0.6;\) and \(a = 0.6, b = 1\) (Fig. 3). With this asymmetry in the signal, it is clear that asymmetric CPCCA-1 solutions show the highest skill. For \(a > b\) \((b > a)\), the highest values for FSVE, COR, and SCF tend to result from choosing \(a > b\) \((b > a)\). It is plausible that as the signal gets smaller, variance/covariance becomes less of a distinguishing property, and lowering its role in constraining solutions consequently leads to higher skill. This systematic sensitivity illustrates a potential caveat if asymmetric solutions (including RDA) are based on an incorrect expectation of the signal-to-noise ratio. However, if an asymmetry is reasonably expected, asymmetric CPCCA-1 solutions can provide some moderate benefit in isolating the signal as a whole.

**An objective parameter choice for regularization**

Without prior knowledge, a more objective basis is required for choosing \(a\) and \(b\) before applying CPCCA as a regularization. In the synthetic experiments, solutions with higher (lower) cross-validated COR tend to lie closer to (farther from) the optimal solution on average, and typically cross validation is an effective approach for quantifying overfitting. However, in practical cases, there may not be enough data for both determining the best parameter values and applying CPCCA for some purpose (as doing both on the same piece of data would artificially inflate skill). For this reason, an objective parameter choice is proposed here based on Ledoit and Wolf (2004), who developed an optimal and objective choice for the ridge parameter, which may be applied to RCCA (Cruz-Cano and Lee 2014).

The ridge regression transformation (when applied to whitening) involves adding a small positive value to the diagonal of the sample covariance matrix prior to taking
the inverse square root, which may be represented for $X$ by the transformation

$$
T_{\text{ridge}} = [(1 - \rho)XX^T + \rho \mu I]^{-1/2},
$$

(19)

with constant $\mu = \|X\|_F^2/m$ such that the total variance is maintained. In this particular form, varying $\rho$ from zero to one results in a path of linear transition that is directly comparable to that of partial whitening, given in Eq. (1). However, an objective and optimal choice for the ridge parameter $\rho$ is desirable for practical applications. Regarding the estimation of an unknown population covariance matrix, Ledoit and Wolf (2004) elegantly derive a choice for $\rho$ that is asymptotically optimal (as both the sample size $n$ and number of spatial points $m$ increase together) and objective in that it is determined from the sample data $XX^T$ alone. This is given by

$$
\rho_{\text{LW}} = \frac{1}{n^2} \sum_{i=1}^n (n-1)x_i^T xx^T - \mu ||x_i||_F^2
$$

(20)

Provided that all the synthetic experiments share the same population covariance matrix, $\rho_{\text{LW}}$ is computed with Eq. (20) using $n = 30$ and averaged across 100 synthetic experiments to yield a value of about 0.19.

To the author’s knowledge, no optimal and objective parameter choice for continuum power regression (and consequently CPCCA) has been developed, and a derivation analogous to that made by Ledoit and Wolf (2004) for the ridge parameter is either highly complex or does not exist. Alternatively, here a solution is chosen that estimates the sample covariance matrix closest to the Ledoit–Wolf estimate. This is found iteratively by varying $\alpha$ to minimize mean-square error, given by

$$
\min ||\nu XX^T(1-\alpha) - \nu XX^T - \rho_{\text{LW}} \mu I||_F^2,
$$

(21)

with $\nu = \|X\|_F^2/\|XX^T(1-\alpha)\|_F^2$ so that the total variance is maintained. On average, this yields a parameter value ($\alpha$) of about 0.20.

Interestingly, this Ledoit–Wolf-based estimate that uses partial whitening is consistently closer to the population covariance than the Ledoit–Wolf estimate itself (both of which only have knowledge of the population covariance from the particular sample). The Ledoit–Wolf estimate yields on average a 7% reduction in the mean-square error between the population covariance and sample covariance $XX^T$, whereas the closest partial whitening estimate yields an average reduction of 17% with less error in over 90% of the realizations. This difference is reduced but still evident if $X$ consists of only background noise with a spatial autocorrelation of 0.85 or greater (not shown). Thus, shifting the Ledoit–Wolf covariance estimate to the closest estimate using partial whitening can yield significant improvement. This is a strong result considering that the resulting estimate is presumably suboptimal (in terms of $\alpha$).

After determining $\alpha$ and $\beta$ using the Ledoit–Wolf-based estimate, CPCCA is computed and results are contrasted with RCCA using the Ledoit–Wolf estimate (Cruz-Cano and Lee 2014) in terms of isolating the coupled signal (Table 1). Interestingly, RCCA-1 is consistently and appreciably more skillful in isolating the signal, something counterintuitive considering the relative skill in the sample covariance estimates. This, however, is likely related to RCCA acting as a more effective filter for weaker EOFs that are dominated by noise. After taking the approach of Barnett and Preisendorfer (1987) and truncating $X$ and $Y$ to the leading 10 EOFs, CPCCA-1 consequently has higher skill overall and compared with RCCA-1 (Table 1). Note also that the Ledoit–Wolf-based estimate that uses partial whitening remains closer to the population covariance than the Ledoit–Wolf estimate. It is possible that such results may better
resemble typical situations in which coupled modes more broadly span PC space. This at least suggests that CPCCA is comparable to RCCA in terms of isolating a coupled mode for a range of signal-to-noise ratios, thus confirming that CPCCA is an effective regularization for CCA. Also, fitting the transformed sample covariance to the Ledoit–Wolf covariance estimate provides an objective basis for choosing CPCCA parameter values.

5. Summary and conclusions

Given the various approaches for analyzing covariance and isolating linear relationships between multiple datasets, regularization and unified (or continuum) approaches are beneficial for reducing the effect of sample noise. Ridge regression has been used to regularize CCA (RCCA; Vinod 1976); although up to this point the comparable technique of continuum power regression (Lorber et al. 1987; Stone and Brooks 1990) has not yet been similarly applied. This is done in this study for which continuum power CCA (CPCCA) is introduced as a new and effective approach. As opposed to applying the full whitening transformation to individual variables during the computation of CCA, CPCCA applies the partial whitening transformation. The partial whitening transformation acts to decorrelate variables, equivalently removing disparity in EOF weighting but only to a partial degree that is specified according parameters \(\alpha\) and \(\beta\). This acts to condition the sample covariance through exponential shrinking and can provide an alternative estimate that is closer to the population covariance. CPCCA maximizes the squared cross covariance between partially whitened variables. As a unified approach, the range of CPCCA solutions bridge those of CCA \((\alpha = \beta = 0)\), MCA \((\alpha = \beta = 1)\), and RDA \((\alpha = 0\) and \(\beta = 1\) or vice versa) with PCR as an upper limit \((\alpha \to \infty\) or \(\beta \to \infty)\).

Along the \(\alpha-\beta\) plane, the recommended symmetric CPCCA solutions are a regularization for CCA (typically small \(\alpha\) and \(\beta\)), a variance bias correction for MCA \((\alpha = \beta = 0.5)\), and a regularization for RDA (typically small \(\alpha\) with \(\beta = 1\): all of which are continuously linked by CPCCA. In the idealized synthetic experiments, solutions using fractional values for \(\alpha\) and \(\beta\) tend to produce the strongest relationships with relatively higher skill in isolating a known coupled mode compared with traditional solutions. This holds true for both the undersampled and well-sampled synthetic relationships tested in this study (not shown).

A second-order finding from the synthetic experiments is that parameter values associated with the highest skill tend to be proportional to the actual signal-to-noise ratio. It is suggested that the effectiveness of constraining CPCCA according to variance/covariance is dependent on the magnitude of the signal. This result and interpretation sheds light on the expected differences between CCA, MCA, and RDA solutions. It appears to be of most consequence for situations involving asymmetry in the signal-to-noise ratio, in which case asymmetric solutions are systematically more skillful.

Just as ridge regression and continuum power regression provide distinct continuous paths between different approaches for linear regression, RCCA and CPCCA provide two different planes of solutions that bridge CCA, MCA, and RDA. RCCA takes a linear transition, whereas CPCCA takes an exponential transition. It is unclear which plane is preferable for isolating coupled modes, and it likely depends on the data. However, in terms of estimating the population covariance from sample data alone, in the synthetic experiments conducted here, partial whitening tends to provide CPCCA with a significantly more accurate covariance estimate than ridge regression does for RCCA. This contrast is found to be strongly related to the autocorrelation of the data (among spatial points or predictors) and, depending on aspects of the coupled signal, this may or may not lead to a more accurate representation of that signal.

Further work should be done to apply CPCCA to observed physically meaningful relationships: for example, variability of the physical climate system. The synthetic experiments in section 4 intend to mimic such physical relationships; however, only one true coupled signal is present with the correct assumption that remaining variability is independent noise. This assumption is, however, invalid in the real physical climate system [given multiple nonorthogonal modes such as ENSO and the North Atlantic Oscillation (NAO)], and the ability of CPCCA to isolate multiple coupled modes in an idealized systematic framework remains to be tested. It is unclear whether differing orthogonality constraints across the range of
CPCCA solutions have an additional impact on the ability to isolate multiple coupled modes, although it is hypothesized that the relative skill is similar between trailing solutions. It is plausible that better representing a dominant relationship could lead to better representing a weaker independent relationship. This is consistent with recent work applying CPCCA to isolate observed tropical–extratropical teleconnections in the climate system, for which CPCCA produces multiple relationships that are quite robust compared with those using other approaches (E. Swenson 2014, unpublished manuscript).

The use of CPCCA is encouraged for applications in multiple other disciplines a range of climate-related applications (as briefly discussed earlier): namely, diagnosis and prediction, particularly paleoclimate reconstruction and seasonal statistical downscaling. In this direction, further work is underway to apply CPCCA to downscale land temperature and precipitation in efforts to correct bias in dynamical models. CPCCA is also encouraged for a range of climate applications as well as applications in other disciplines.

Acknowledgments. This work could not have been completed without discussion, advice, and encouragement from T. DelSole. The author would also like to thank D. M. Straus, J. H. Yoo, and J. Shukla for valuable discussion and input. This work was supported by the APEC Climate Center.

APPENDIX A

CPCCA in PC Space for Small Samples

a. Principal component analysis

Principal component analysis (PCA) is a useful technique for data compression that decomposes a variable into a unique orthogonal basis set with leading modes that maximize variance in a consecutive manner. As noted by Wilks (2011), PCA may be computed using singular value decomposition (SVD). Given X and Y (as described in section 2), assume also now that the number of spatial points is more than the number of time samples (m × n). SVD results in $X = USV_x^T$ and $Y = USV_y^T$, from which the weighted pattern vectors (EOF patterns) are given simply by $U_x$, $U_y$, and $V_x$, $V_y$ with associated variates (PCs) given by $S_x$ and $S_y$, respectively.\(^{a1}\) In this case, an economy SVD may be computed such that $U_x$ is not square with dimensions of $m_x \times n$ such that $U_x^T U_x = I$ but $U_x^T U_x^T \neq I$. Then, $S_x$ and $V_x$ are both square matrices with dimensions of $n \times n$. Matrix $V_x$ is unitary and $S_x$ is a diagonal matrix with diagonal elements giving the positive sorted PC singular values that represent individual standard deviations of the corresponding EOFs. Properties hold for $U_y$, $S_y$, and $V_y$ generated from PCA computed for $Y$.

b. Partial whitening in terms of EOFs and PCs

After computing PCA of $X$ and $Y$ with $X = U_xS_xV_x^T$ and $Y = U_yS_yV_y^T$, respectively, individual covariance matrices may be expressed in terms of EOFs as $XX^T = U_xS_x^2U_x^T$ and $YY^T = U_yS_y^2U_y^T$, respectively. Covariance matrices are assumed to be of rank $n$. Thus, when computing the whitening transformation, matrices are singular and thereby invertible when $m_x > n$ and $m_y > n$. In this case, the Moore–Penrose pseudoinverse may be used instead, indicated by $(\cdot)^\dagger$. Then the whitening transformation may then be expressed by $(XX^T)^{(1-\alpha)/2} = (U_xS_x^\alpha U_x^\dagger)^\dagger = U_xS_x^{\alpha - 1}U_x^T$ and similarly for $Y$. From this it may be demonstrated that whitening simply acts to weight EOFs equally by replacing $S_x$ with the identity matrix, with $U_xS_x^{-1}U_x^T X = U_xV_x^T$. The partial whitening transformation of Eq. (1) may be similarly expressed as

$$(XX^T)^{(1-\alpha)/2} = (U_xS_x^\alpha U_x^\dagger)^\dagger = U_xS_x^{\alpha - 1}U_x^T$$

Analogously, partial whitening preserves EOF patterns and variates while applying PC singular values to the powers of $\alpha$ and $\beta$, yielding $X^\alpha = T_xX = U_xS_x^\alpha V_x^T$ and $Y^\beta = T_yY = U_yS_y^\beta V_y^T$, respectively. Equivalently, partially whitened variables only differ in terms of EOF standard deviations $S_x^\alpha$ and $S_y^\beta$. As $\alpha$ and $\beta$ are decreased from 1 to 0, the disparity among EOF magnitudes is similarly decreased until there is none. In PC space, partially whitened variables are represented by the transformed weighted PCs, which are recovered by projecting $X^\alpha$ and $Y^\beta$ onto the unweighted orthogonal EOF patterns, yielding $X^\alpha U_x = V_x S_x^\alpha$ and $Y^\beta U_y = V_y S_y^\beta$. Then the cross covariance may be expressed as $U_x^\dagger X^\alpha Y^\beta U_y = S_x^\alpha V_x^\dagger V_y S_y^\beta$, an $n \times n$ matrix.

c. Continuum power CCA in PC space

In PC space, the cross covariance between $X$ and $Y$ may be represented in terms of only the PCs with $U_x^\dagger XY^\beta U_y = S_x V_x^\dagger V_y S_y$. Then the CPCCA constrained optimization problem of Eqs. (2) and (6) can be re-expressed as

\(^{a1}\)For $X$, note that PCA is classically computed using an eigenvector decomposition of the covariance matrix from which pattern vectors result directly while variates must be obtained by projection. The use of SVD eliminates the need to both construct the covariance matrix and project the data.
max \( q_i^T S_{xy} V_j^T V_j y_j q_j \) \( \text{(A2)} \)

subject to

\[ q_i^T S_{xx}^{2(1-\alpha)} q_x = 1 \text{ and } q_j^T S_{yy}^{2(1-\beta)} q_y = 1. \] \( \text{(A3)} \)

If one substitutes appropriate values for \( \alpha \) and \( \beta \), it is clear that the only distinction between CCA-1 (\( \alpha = \beta = 0 \)), MCA-1 (\( \alpha = \beta = 1 \)), and RDA-1 (\( \alpha = 0 \) and \( \beta = 1 \)) is the role of the PC singular values in the norm constraints, given in Eq. (A3). Following Eqs. (7) and (8), Eqs. (A2) and (A3) may also be expressed as MCA-1 involving partially whitened weighted PCs, which involves computing SVD of the cross-covariance with

\[ S_{xy} V_j^T V_j y_j = U' S' V'^T. \] \( \text{(A4)} \)

Note that computing SVD in PC space is generally highly desirable if spatial dimensions greatly exceed the number of time samples. This is because the computational time of computing SVD of the \( n \times n \) matrix cross-covariance matrix, expressed by Eq. (A4), in addition to that taken for computing individual PCA, can be much less than the time taken for computing SVD of an \( m_x \times m_y \) matrix, as expressed by Eq. (9). Unitary matrices \( U' \) and \( V' \) provide linear combinations of the partially whitened weighted PCs corresponding to individual CPCCA modes, and they are equivalent to those in Eq. (9) after a transformation back to physical space, with \( U = U_x U' \) and \( V = U_y V' \). Solutions to Eqs. (A2) and (A3) may then be given as \( q_i = S_{xx}^{a-1} u_i' \) and \( q_j = S_{yy}^{b-1} v_j' \) which produce the highest possible cross-covariance value of \( s'_i \).

The physical patterns for CPCCA-1 are recovered by transforming solutions back to physical space, and the full set of CPCCA patterns is given by

\[ P_x = U_x S_{xx}^{a-1} U' \text{ and } P_y = U_y S_{yy}^{b-1} V'. \] \( \text{(A5)} \)

The associated CPCCA variates are found by taking linear combinations \( U' \) and \( V' \) of the partially whitened weighted PCs given by

\[ R_x = V_x S_{xx}^{a} U' \text{ and } R_y = V_y S_{yy}^{b} V'. \] \( \text{(A6)} \)

For \( X \), the orthogonality conditions of Eqs. (12) and (13) may be expressed in terms of EOFs and PCs with

\[ P_x^T P_x = U'^T S_{xx}^{2(a-1)} U' = I \text{ when } \alpha = 1 \text{ and } (A7) \]

\[ R_x^T R_x = U'^T S_{xx}^{2a} U' = I \text{ when } \alpha = 0. \] \( \text{(A8)} \)

\( d. \) Upper limit of continuum power CCA

As parameters get large, PC singular values of partially whitened variables diverge. For \( X \), after scaling by the leading value \( s_{x,1}^a \), this is evident in that

\[ \lim_{a \to \infty} S_{x,i}^a = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}. \] \( \text{(A9)} \)

Provided that weighted PC-1 of \( X \) has nonzero covariance with the PCs of \( Y \), Eq. (A9) also means that PC-1 will produce the highest cross covariance and, as \( \alpha \to \infty \), SVD from Eq. (A4) yields \( u_i' \to (1 \ 0 \ \cdots \ 0)^T \). After substituting into Eqs. (A5) and (A6), this results in \( p_{x,1} \to \pm S_{x,1}^{-1/a} u_{x,1} \) and \( r_{y,1} \to \pm S_{y,1}^{b-1} v_{y,1} \) recovering EOF-1 and PC-1, respectively, after normalization. Then CPCCA-1 for \( Y \) may be found through linear regression with PC-1 for \( X \). It may be similarly demonstrated that trailing modes converge to trailing EOFs, and thus CPCCA approaches PCR as \( \alpha \to \infty \).

**APPENDIX B**

**Relationship with Linear Regression**

The sample covariance matrix between partially whitened variables (using fixed parameters), given by \( XX^T \), also represents the transpose of a matrix of linear regression coefficients involving transformed data. Without any transformation, the least squares solution between predictor \( x \) and predictand \( y \) (time-varying column vectors of \( X \) and \( Y \), respectively) is given by \( y_{is} = Ax \), with the regression matrix \( A \) given by

\[ A = YY^T (XX^T)^{-1}. \] \( \text{(B1)} \)

Similarly, transformed variables \( \tilde{X} \) and \( \tilde{Y} \) may be formed with solution \( \tilde{y}_{is} = \tilde{A} \tilde{x} \) such that \( \tilde{A} = YY^T (XX^T)^{-1} = Y^T X^T \). As in Tippett et al. (2008), who express cross covariance as a regression matrix, this is satisfied with \( \tilde{X} = (XX^T)^{-1} X \) and \( \tilde{Y} = Y^* \). In terms of the original variables and their EOFs and PCs, this reduces to

\[ \tilde{X} = T_x^{-1} (XX^T)^{-1} X = (XX^T)^{(-\alpha - 1)/2} X = U_x S_x^{a-1} V_x^T \] \( \text{(B2)} \)

\[ \tilde{Y} = T_y^{-1} Y (YY^T)^{(\beta-1)/2} Y = U_y S_y^{b} V_y^T. \] \( \text{(B3)} \)

In terms of \( \tilde{A} \), note that \( \tilde{A} = T_y^{-1} A T_x^{-1} (XX^T)^{-1} \). After substituting \( \tilde{X} = P_x R_x^T \) and \( \tilde{Y} = P_y R_y^T \), into Eq. (B1), the
original regression matrix may be simplified to a form involving the CPCCA variates with
\[
AP_x = P_y R^T_x R_x (R^T_x R_x)^{-1}.
\]
(B4)

In practice, typically a limited number of the leading CPCCA modes are used for linear regression for the purpose of constructing a prediction model. This involves replacing the full sets of patterns and variates in Eq. (B4) with truncated representations and computing a Moore–Penrose pseudoinverse rather than the full inverse (which is necessary in the first place if $XX^T$ is invertible).

**APPENDIX C**

**Synthetic Data**

The 10 population EOF patterns ($E^{\text{pop}}$) for both $X$ and $Y$ (both of spatial dimension $m = 60$) for all experiments are given by

\[
E_{ij}^{\text{pop}} = \begin{cases} 
\sin \left( \frac{i-1}{m} \pi j \right) , & \text{if } j = 1, 3, 5, 7, 9 \\
\cos \left( \frac{i-1}{m} (j-1) \pi \right) , & \text{if } j = 2, 4, 6, 8, 10 
\end{cases}
\]

where $E_{ij}^{\text{pop}}$ is the $i$th row and $j$th column of $E^{\text{pop}}$. The associated variances are given by diagonal elements of $(E^{\text{pop}})^TE^{\text{pop}} = S^2$, a diagonal $10 \times 10$ matrix. The corresponding population PCs, or columns of $F_x^{\text{pop}}$ and $F_y^{\text{pop}}$, are randomly generated yet constrained such that specified linear combinations $U^{\text{pop}}$ and $V^{\text{pop}}$ yield variates $R^{\text{pop}}$ and $S^{\text{pop}}$ with expected cross correlation $C^{\text{pop}}$ given by

\[
\langle (U^{\text{pop}})^T (F_x^{\text{pop}})^T F_y^{\text{pop}} V^{\text{pop}} \rangle = \langle (R_x^{\text{pop}})^T R_y^{\text{pop}} \rangle = C^{\text{pop}},
\]

(C2)

where $\langle \mathbf{B} \rangle$ denotes the expectation of $\mathbf{B}$. For all experiments, diagonal elements of $C^{\text{pop}}$ are specified as

\[
C_{ij}^{\text{pop}} = \begin{cases} 
0.9 , & \text{if } i = 1 \\
0 , & \text{if } i > 1 
\end{cases}
\]

such that only one true coupled mode is present with a correlation of 0.9. The variates are generated randomly by first computing a Cholesky decomposition of the $20 \times 20$ correlation matrix, given by

\[
C = \begin{bmatrix} 
I & C^{\text{pop}} \\
(C^{\text{pop}})^T & I 
\end{bmatrix} = U \Sigma U^T.
\]

and then multiplying a randomly generated $n \times 20$ white noise matrix $W^{\text{pop}}$ by $U$. With the white noise such that $\langle U_{ij} (W^{\text{pop}})^T W^{\text{pop}} U_{ij} \rangle = I$ and $\langle (W^{\text{pop}})^T W^{\text{pop}} \rangle = U_U U^T = C$, variates satisfying Eq. (C2) may be generated from the columns of $W^{\text{pop}} U = (R_x^{\text{pop}} - R_y^{\text{pop}})$. Provided with unitary matrices $U^{\text{pop}}$ and $V^{\text{pop}}$, population PCs are then produced from $F_x^{\text{pop}} = R_x^{\text{pop}} U^{\text{pop}}$ and $F_y^{\text{pop}} = R_y^{\text{pop}} V^{\text{pop}}$, with $\langle (F_x^{\text{pop}})^T F_x^{\text{pop}} \rangle = I$ and $\langle (F_y^{\text{pop}})^T F_y^{\text{pop}} \rangle = I$. $U^{\text{pop}}$ and $V^{\text{pop}}$ are also random, with exception to the weighting of population EOF-1, which is specified as $U_{11}^{\text{pop}} = \sqrt{a}$ and $V_{11}^{\text{pop}} = \sqrt{b}$ with $0 \leq a \leq 1$ and $0 \leq b \leq 1$ for $X$ and $Y$, respectively.

Given that only one signal is present in $X$ and $Y$ with temporal variation given by $r_{x,1}^{\text{pop}} = F_{x,1}^{\text{pop}} U_{11}^{\text{pop}}$ and $r_{y,1}^{\text{pop}} = F_{y,1}^{\text{pop}} V_{11}^{\text{pop}}$, the associated signal pattern vectors are given by $p_x^{\text{signal}} = E_{x,1}^{\text{pop}} U_{11}^{\text{pop}}$ and $p_y^{\text{signal}} = E_{y,1}^{\text{pop}} V_{11}^{\text{pop}}$, respectively. The signal may then be expressed as $X^{\text{signal}} = p_x^{\text{signal}} (r_{x,1}^{\text{pop}})^T$ and $Y^{\text{signal}} = p_y^{\text{signal}} (r_{y,1}^{\text{pop}})^T$, respectively. The trailing columns of $U^{\text{pop}}$ and $V^{\text{pop}}$ provide EOF linear combinations corresponding to the subspace orthogonal to the signal, essentially background noise that fills the space spanned by the population EOFs, given by $E^{\text{pop}} (F_x^{\text{pop}})^T X^{\text{signal}}$ and $E^{\text{pop}} (F_y^{\text{pop}})^T Y^{\text{signal}}$, respectively. By design, had the population PCs been strictly orthonormal across the sample (not only in the expectation), CCA-1 between $E^{\text{pop}} (F_x^{\text{pop}})^T$ and $E^{\text{pop}} (F_y^{\text{pop}})^T$ would recover the signal exactly.

Given EOF-1 weightings $a$ and $b$, the expectation of the true signal variance depends on the variance of population EOF-1 ($s_x^2$, the first diagonal element of $S$) and the sum of variances of the remaining population EOFs, which is weighted evenly by $1 - a$ and $1 - b$, respectively. For all experiments, $s_x^2 = 0.25$ and the sum of variances of the first 10 EOFs is 0.8. Including additional background noise, the total expected variance is specified to be equal to 1 (demonstrated later). The expected signal variance may then be given by

\[
\langle \| X^{\text{signal}} \|_F^2 \rangle = \langle \| p_x^{\text{signal}} \|_F^2 \rangle = 0.25a + 0.55(1 - a)/9 \\
\langle \| Y^{\text{signal}} \|_F^2 \rangle = \langle \| p_y^{\text{signal}} \|_F^2 \rangle = 0.25b + 0.55(1 - b)/9.
\]

(C5)

Increasing $a$ or $b$ from 0 to 1 increases the expected true signal variances from $0.55/9 = 0.061$ to 0.25, corresponding to a signal-to-noise ratio from $[0.061/(1 - 0.061)]^{1/2} \approx 0.26$ to $[0.25/(1 - 0.25)]^{1/2} \approx 0.58$, respectively. Also, the disparity of the individual variances decays exponentially according to $s_i^2 = 0.25 \tau^{-i}$, with $\tau \approx 0.696$ chosen so that the sum is 0.8.

The time-varying population EOFs are embedded in randomly generated background noise $N$ with spatial
(temporal) variation following red (white) noise and with expected variance of $s^2_{\text{noise}} = 0.2$. Similar to Bretherton et al. (1992), red noise is generated by a Markov process given by

$$n_i = r n_{i-1} + \sqrt{1-r^2} w_i,$$  \hspace{1cm} (C6)

where $n_i$ and $w_i$ are the $i$th rows of $N$ and $W$, respectively, both $m \times n$ matrices. Equation (C6) states that the $i$th time-varying spatial point of the background noise $n_i$ is given by a weighted average between the value at its neighboring spatial point $n_{i-1}$ and random white noise $w_i$, such that the noise has unit variance at each spatial point with an autocorrelation of $r$. Given that $\langle w_i^j w_i^j \rangle = 1$ and $\langle n_i^j n_i^j \rangle = 0$ for $i \neq j$, with Eq. (C6) this is demonstrated in that $(n_i^j n_i^j) = 1$ and $(n_i^j n_{i-1}^j) = r$ in all experiments, $r = 0.9$, and the background noise is independently generated in the same manner for both $X(N_i)$ and $Y(N_i)$. Finally, the total randomly generated datasets are given by

$$X = E^{\text{pop}}(F^{\text{pop}}_x)^T + s_{\text{noise}} N_x \text{ and}$$

$$Y = E^{\text{pop}}(F^{\text{pop}}_y)^T + s_{\text{noise}} N_y,$$  \hspace{1cm} (C7)

For $X$, given that $\langle |x_i|^2 \rangle = m$ and $\langle |x_i F^{\text{pop}}(E^{\text{pop}})^T | \rangle \rangle$ = 0, with substitution of Eq. (C7) it may be shown that $\langle |x_i|^2 \rangle = tr(S_0^2) + s_{\text{noise}}^2 = 1$, with $tr(B)$ indicating the trace or sum of diagonal elements of $B$. Keep in mind that the “noise” in the signal-to-noise ratio refers to all of the variance that does not contribute to the signal (including that spanned by the population EOFs), and it is not equivalent to $s_{\text{noise}}^2$ given here. Along with sampling variability of the population EOFs, the background noise randomly projects onto the population EOFs, producing nonzero covariance and cross covariance, leading to additional overfitting when computing CPCCA.

REFERENCES


