Environmental Modelling & Software 26 (2011) 1363–1371

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Short communication

Statistical downscaling of rainfall data using sparse variable selection methods

A. Phatak [a,b,*], B.C. Bates [a,c], S.P. Charles [a,d]

[a] CSIRO Climate Adaptation Flagship, Australia
[b] CSIRO Mathematical & Information Sciences, Private Bag 5, Wembley, WA 6913, Australia
[c] CSIRO Marine & Atmospheric Research, Private Bag 5, Wembley, WA 6913, Australia
[d] CSIRO Land & Water, Private Bag 5, Wembley, WA 6913, Australia

ARTICLE INFO

Article history:
Received 15 September 2009
Received in revised form
4 May 2011
Accepted 8 May 2011
Available online 14 June 2011

Keywords:
Statistical downscaling
Variable selection
*L*1-norm
Sparse variable selection
Logistic regression

ABSTRACT

In many statistical downscaling methods, atmospheric variables are chosen by using a combination of expert knowledge with empirical measures such as correlations and partial correlations. In this short communication, we describe the use of a fast, sparse variable selection method, known as RaVE, for selecting atmospheric predictors, and illustrate its use on rainfall occurrence at stations in South Australia. We show that RaVE generates parsimonious models that are both sensible and interpretable, and whose results compare favourably to those obtained by a non-homogeneous hidden Markov model (Hughes et al., 1999).

Crown Copyright © 2011 Published by Elsevier Ltd. All rights reserved.

1. Introduction

General circulation models (GCMs) constitute the most important set of tools for encoding our understanding of the mechanisms that govern large-scale climate patterns (Murphy et al., 2004). First, they provide important evidence towards the conclusion that increases in anthropogenic forcing are responsible for the observed increase in globally averaged temperatures since the mid-20th century (Solomon et al., 2007 §9.4.1.2). Second, they are used to generate scenarios describing possible future climates, information that is needed to facilitate adaptation and mitigation of the effects of climate change. Nevertheless, for local impacts studies, the spatial resolution of GCMs is still too coarse, hence the widespread use of downscaling methods to bridge the gap between synoptic-scale information from GCMs and the spatially and temporally fine scale information such as daily rainfall required in, for example, hydrological studies.

Downscaling methods are usually classified into two groups: dynamical downscaling and statistical downscaling (SD) (Wilby and Wigley, 1997). Dynamical downscaling relies on the use of regional climate models to produce outputs at higher resolutions. By contrast, SD methods construct *empirical* models that relate a small set of coarse-scale predictor variables to local climate variables—the predictands—such as temperature or rainfall occurrence and amounts. There are many different types of statistical downscaling methods (Fowler et al., 2007). They are all based on the implicit assumption that (a) the predictors that have been chosen, and their relationships to the predictands, are physically meaningful; and (b) these relationships will hold in the future in a changed climate.

The *choice* of predictors is also crucial but, as Fowler et al. (2007) point out, there is often "little consensus on the most appropriate choice ...." The recent review of methods for precipitation downscaling by Maraun et al. (2010) suggests some desirable properties of predictors for statistical downscaling: that they be informative, well simulated by dynamical models, and capture the effects of climate change. In typical downscaling applications, the set of potential predictors comprises atmospheric variables obtained over a grid. Hence, given the ensemble of atmospheric variables and the grid identified by the investigator as being potentially relevant for downscaling, the selection problem involves choosing a subset of atmospheric variables and their grid location(s) for use in a downscaling model. For the most part, the selection of predictors in any given downscaling application combines expert knowledge with measures such as correlations and partial correlations between the predictors and predictand, and often involves iteratively fitting

* Corresponding author. CSIRO Mathematical & Information Sciences, Private Bag 5, Wembley, WA 6913, Australia. Tel.: +61 8 9333 6184; fax: +61 8 9333 6121.
E-mail addresses: Aloke.Phatak@csiro.au (A. Phatak), Bryson.Bates@csiro.au (B.C. Bates), Steve.Charles@csiro.au (S.P. Charles).

1364-8152/$ – see front matter Crown Copyright © 2011 Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2011.05.007

models with different sets of predictors to find the most useful one (e.g. Hughes et al., 1999; Chandler and Wheater, 2002; Wilby et al., 2002; Chandler, 2005). Using such methods, however, it is only possible to consider a relatively small subset of predictors. More recently, Hessami et al. (2008) described an automatic variable selection method based on backward stepwise regression, but they too limited the ensemble of predictors to about 30, from which five were selected. Other methods for selecting small subsets of variables include genetic algorithms (Traveria et al., 2010) and the semi-automatic method of Prasad et al. (2010), which consists of ranking individual variables and then using all-subsets selection.

An alternative to selection of individual variables from a spatial field is to employ derived variables such as principal components (PCs) or canonical variates (for example, Schoof et al. (2009)). These derived variables can also be used for dimension reduction and then in regression (PCs) or canonical correlation analysis, and they may, in some instances, provide a meaningful physical interpretation and identify modes of variability (though not always—see, for example, the caveats in Zwiers and Von Storch (2004) and Hannachi et al. (2007)).

In this short communication, we describe the use of a fast, sparse variable selection method for selecting atmospheric predictors in a logistic regression model for downscaling daily rainfall occurrence at stations in South Australia. The method, known colloquially as RaVE (for Rapid Variable Elimination), is due to Kiiveri (2008), and although it is related to implicit variable selection methods such as LASSO—least absolute shrinkage and selection operator (Tibshirani, 1996)—it provides a much more flexible framework for model fitting and variable selection. We illustrate the method by predicting daily rainfall occurrence from a set of 392 potential predictors (7 variables at 56($= 8 \times 7$) grid points), and show that it generates parsimonious models that are both sensible and interpretable and whose results compare favourably to those obtained by a non-homogeneous hidden Markov model (NHMM) (Hughes et al., 1999; Charles et al., 2004). We also note that RaVE can be used to select which PCs to use in, for example, a principal component regression.

2. Logistic regression and sparse variable selection

2.1. Logistic regression

Logistic regression is one of a large class of *generalized linear models* (McCullagh and Nelder, 1989), and is often used to model the probability of rainfall occurrence as a function of predictors in statistical downscaling applications (e.g. Chandler and Wheater, 2002; Furrer and Katz, 2007; Hessami et al., 2008). When the presence or absence of rainfall in the previous q time intervals is included as a covariate, logistic regression can be used to fit a binary Markov chain of order q. Moreover, by including additional predictors such as atmospheric variables from reanalysis data or GCM outputs, a non-homogenous Markov chain can be fitted (Fahrmeir et al., 2001, Chapter 6), and hence the transition probabilities will be modulated by the outputs included in the model.

We denote rainfall occurrence as Y, and following convention, let $Y = 1$ if it rained on a given day, $Y = 0$ if it did not. Then, the binary logistic regression model can be written as

$$\log\left[\frac{p(Y=1|\boldsymbol{x})}{1-p(Y=1|\boldsymbol{x})}\right] = \text{logit}(p_Y) = \boldsymbol{x}^T\boldsymbol{\beta} \tag{1}$$

where $\boldsymbol{x} = (1, x_1, x_2, \ldots, x_p)^T$ is a $(p+1)\times 1$ vector of covariates $x_j, j = 1, 2, \ldots, p$, which may include, for example, atmospheric variables such as geopotential height, specific humidity, and dew-point temperature depression at different pressure levels, mean sea-level pressure, but also previous values of rainfall $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-q}$; $p(Y=1|\boldsymbol{x})$ is the probability of rain given the covariates included in the model; and $\boldsymbol{\beta}$ is a $(p+1)\times 1$ vector of coefficients to be estimated.

2.2. Shrinkage, or regularization, methods

When the number of covariates is potentially very large and they are highly correlated, estimation methods such as least squares can lead to predictions that have a large variance. As Hastie et al. (2001, p. 55) point out, however, "prediction accuracy can sometimes be improved by shrinking or setting some coefficients to zero." Shrinkage, or regularization, means reducing the size of the coefficients so that they are 'pulled' towards zero, whereas setting some coefficients to zero implies variable selection. Examples of commonly used regularization methods include ridge regression and principal component regression, and methods that combine *both* shrinkage and variable selection include LASSO (Tibshirani, 1996) and RaVE (Kiiveri, 2008). In addition to reducing variance, these methods also provide information that assists in interpreting the results: principal components, for example, that provide a visual summary of the data, or a reduced set of variables that exhibit the strongest effect on the predictand of interest.

The vector of coefficients in eq. (1) is usually estimated by maximum likelihood; in the case of LASSO, the log-likelihood is penalized by the addition of an $L1$ penalty term of the form $\lambda\Sigma_{i=1}^{p}|\beta_i|$, where $\lambda > 0$ is known as the shrinkage, or regularization, parameter which controls both variable selection and the amount of shrinkage. The LASSO solution can also be interpreted as the Bayes posterior mode under independent double-exponential (Laplace) priors for the coefficients (Tibshirani, 1996), i.e., $p(\beta_i) \propto \exp(-\lambda|\beta_i|)$; hence, the $L1$ penalty in the penalized likelihood corresponds to the negative log-prior.

2.3. Rapid variable elimination (RaVE)

The approach we adopt here, which is due to Kiiveri (2008), can also be interpreted in the penalized likelihood and Bayesian frameworks. However, by contrast with the original formulation of LASSO (Tibshirani, 1996), the prior distribution of the coefficients—and hence the penalty term—is derived in an explicitly hierarchical manner. It reflects the assumption that most of the elements of $\boldsymbol{\beta}$ may be zero, or at least very small, and it can be specified in two parts as

$$\beta_i\big|\nu^2 \sim \text{N}\left(0, \nu^2\right), \quad \nu^2 \sim \text{Ga}(k, b) \tag{2}$$

This hierarchical prior is known as a 'normal-gamma' (NG) prior. In the first part of the specification, the ith coefficient has a normal distribution with mean zero and variance ν^2; in the second, the variance ν^2 has a gamma density with shape k and scale b. The marginal distribution of the β_i can be calculated as $\int_{\nu^2} p(\beta_i|\nu^2)p(\nu^2)\mathrm{d}\nu^2$, and this leads to a prior of the form

$$p(\beta_i) = \left[\frac{2^{(0.5-k)}}{\sqrt{\pi}\Gamma(k)}\right]\frac{\delta K_{(0.5-k)}(\delta|\beta_i|)}{(\delta|\beta_i|)^{(0.5-k)}}, \quad \delta = \sqrt{\frac{2}{b}} \tag{3}$$

where K denotes a modified Bessel function of the third kind (which is a rapidly decaying function), and Γ denotes the gamma function. For fixed values of the hyperparameters k and b (see below), an expectation-maximization algorithm is used to obtain maximum *a posteriori* estimates of the coefficients β_i. Algorithmic details may be found in Kiiveri (2008). Compared to the double-exponential prior in LASSO, the NG prior yields sparser models and gives rise to a wider range of shrinkage behaviour (Griffin and Brown, 2010).

Download English Version:

https://daneshyari.com/en/article/569763

Download Persian Version:

https://daneshyari.com/article/569763

Daneshyari.com