Dealing with many correlated covariates in capture-recapture models

Olivier Gimenez[1], Christophe Barbraud[2]

[1]CEFE UMR 5175, CNRS, Université de Montpellier, Université Paul-Valéry Montpellier, EPHE, 1919 Route de Mende, 34293 Montpellier Cedex 5, France.

[2] CEBC UMR 7372, CNRS – Université de La Rochelle, 79360 Villiers en Bois, France.

Corresponding author: Olivier Gimenez (olivier.gimenez@cefe.cnrs.fr)

Word count: 2412

Summary: Capture-recapture models for estimating demographic parameters allow covariates to be incorporated to better understand population dynamics. However, high-dimensionality and multicollinearity can hamper estimation and inference. Principal component analysis is incorporated within capture-recapture models and used to reduce the number of predictors into uncorrelated synthetic new variables. Principal components are selected by sequentially assessing their statistical significance. We provide an example on seabird survival to illustrate our approach. Our method requires standard statistical tools, which permits an efficient and easy implementation using standard software.

Key words: Animal demography, Population dynamics, Principal-component capture-recapture model, Survival estimation.

INTRODUCTION

Capture-recapture (CR) methods (e.g. Lebreton *et al.* 1992) are widely used for assessing the effect of explanatory variables on demographic parameters such as survival (Pollock 2002). Generally however, complex situations arise where multiple covariates are required to capture patterns in survival. In such situations, one usually favors a multiple regression-like CR modeling framework that is however hampered by two issues: first, because it increases the number of parameters to be estimated, incorporating many covariates results in a loss of statistical power and a decrease in the precision of parameter estimates; second, correlation among the set of predictors – aka multicollinearity – may alter interpretation (see below).

To overcome these two issues, Grosbois *et al.* (2008) recommended to perform a principal component analysis (PCA) on the set of explanatory variables before fitting CR models. PCA is a multivariate technique that explains the variability of a set of variables in terms of a *reduced* set of *uncorrelated* linear combinations of such variables – aka principal components (PCs) – while maximizing the variance (Jolliffe 2002). Grosbois *et al.* (2008) then expressed survival as a function of the PCs that explained most of the variance in the set of original covariates, typically the first one or the first two ones.

However, the main drawback of this approach is that the PCs are selected based on covariates variation pattern alone, regardless of the response variable, and without guarantee that survival is most related to these PCs (Graham 2003). To deal with this issue in the context of logistic regression, Aguilera *et al.* (2006) proposed to test the significance of *all* PCs to decide which ones should be retained, instead of a priori relying on the PCs that explain most of the variation in the covariates.

In this paper, we implement the algorithm proposed by Aguilera *et al.* (2006) to deal with many possibly correlated covariates in CR models, a method we refer to as principal component capture-recapture (P2CR). We apply this new approach to a case study on survival of Snow petrels (*Pagodroma nivea*) that is possibly affected by climatic conditions. In this example, the issue of multicollinearity occurs, and summarizing the set of covariates in a subset of lower dimension is also crucial to get precise survival estimates. Overall, P2CR models can be fitted with statistical programs that perform PCA and CR data analysis. The data and R code are available from GitHub at https://github.com/oliviergimenez/p2cr.

METHODS

We used capture-recapture (CR) models to study open populations over K capture occasions to estimate the probability ϕ_i (i = 1, …, K - 1) that an individual survives to occasion $i+1$ given that it is alive at time i, along with the probability p_j (j = 2, …, K) that an individual is recaptured at time j – aka as the Cormack-Jolly-Seber (CJS) model (Lebreton *et al.* 1992). Covariates were incorporated in survival probabilities using a linear-logistic function:

$$\text{logit}(\phi_i) = \log\left(\frac{\phi_i}{1-\phi_i}\right) = \alpha + \sum_{j=1}^{p} \beta_j X_{ij} \qquad (1)$$

where α is the intercept parameter, X_{ij} is the value of covariate j (j = 1,…, p) in year i (i = 1,…, K - 1), and β_j is its associated slope parameter. Covariates were standardized to avoid numerical instabilities. To assess the significance of a covariate in CR models, we used the analysis of deviance (ANODEV; Skalski, Hoff & Smith 1993) that compares the amount of deviance explained by this covariate with the amount of deviance not explained by this covariate, the CR model with fully time-dependent survival serving as a reference. The ANODEV test statistic is given by:

$$\text{ANODEV} = \frac{\text{Dev}(X) - \text{Dev}(constant)}{1} \Big/ \frac{\text{Dev}(time) - \text{Dev}(X)}{K - 1} \qquad (2)$$

where Dev(*constant*), Dev(*X*) and Dev(*time*) stand for the deviance of models with constant, covariate-dependent and time-dependent survival probabilities. To obtain the associated p-value, the value of the ANODEV is compared with the quantile of Fisher-Snedecor distribution with 1 and K-1 degrees of freedom.

To reduce the dimension of the set of covariates ($X_1, \ldots, X_p$), we used PCA which aims at finding a small number of linear combinations of the original variables – the principal components (PCs) – while maximizing the variance in ($X_1, \ldots, X_p$). Because the variables measurement units often differ, we performed the PCA on the correlation matrix (Jolliffe 2002). To select PCs, we used a forward model selection algorithm as proposed by Aguilera *et al.* (2006) for the logistic regression. The forward algorithm begins with no covariates in the model. Each PC is incorporated in simple linear regression-like CR models and the ANODEV p-value calculated. The PC that has the lowest p-value is added to the null model, say PC_k. Then the PCs that were not retained are incorporated along with PC_k in multiple regression-like CR models, and ANODEV p-values are calculated. In other words, we need to assess the effect of PC_j for $j \neq k$ in the presence of PC_k to decide whether PC_j should be retained. To do so, Dev(*constant*) and Dev(*X*) are replaced by Dev(PC_k) and Dev($PC_k + PC_j$) in Equation 2, where Dev($PC_k + PC_j$) is the deviance of the model with survival as a function of both principal components PC_k and PC_j. We repeat the process until no remaining PC is selected.

All models were fitted using the maximum-likelihood method using MARK (White & Burnham 1999) called with R (Laake 2013).

CASE STUDY

The Snow petrel is a medium sized Procellariiform species endemic to Antarctica that breeds in summer. Birds start to occupy breeding sites in early November, laying occurs in early December and chicks fledge in early March. This highly specialized species only forages within the pack-ice on crustaceans and fishes. Data on survival were obtained from a long-term CR study on Ile des Pétrels, Pointe Géologie Archipelago, Terre Adélie, Antarctica. We refer to Barbraud *et al.* (2000) for more details about data collection. We removed the first capture to limit heterogeneity among individuals, and worked with a total of 604 female capture histories from 1973 to 2002.

The following covariates were included to assess the effect of climatic conditions upon survival variation: sea ice extent (SIE; http://nsidc.org/data/seaice_index/); air temperature, which was obtained from the Météo France weather station at Dumont d'Urville, as a proxy for sea surface temperature; southern Oscillation Index (SOI) as a proxy for the overall climate condition (https://crudata.uea.ac.uk/cru/data/soi/). These environmental variables were averaged over seasonal time periods corresponding to the chick rearing period (January to March: summer period), the non-breeding period (April to June: autumn and July to September: winter), and the laying and incubation period of the same year (October to December: spring). In total, 9 covariates were included in the analysis: sea ice extent in summer (SIEsummer), in autumn (SIEautumn), in winter (SIEwinter), in spring (SIEspring), annual SOI, air temperature in summer (Tsummer), in autumn (Tautumn), in winter (Twinter) and in spring (Tspring).

RESULTS

The CJS model poorly fitted the data (χ^2 = 221.2, df = 127, p << 0.01), and a closer inspection of the results revealed that the lack of fit was explained by a trap-dependence effect (Test2CT, χ^2 = 103.1, df = 27, p << 0.01). Consequently, we estimated two recapture probabilities that differed according to whether or not a recapture occurred the occasion before (Pradel 1993). By first attempting to simplify the structure of recapture probabilities, we were led to consider an additive effect of time and a trap effect (Supplementary material). Estimates of recapture probabilities ranged from 0.14 (standard error [SE] = 0.07) to 0.79 (SE = 0.09) when no recapture occurred the occasion before and from 0.25 (SE = 0.18) to 0.89 (SE = 0.09) when a recapture occurred the occasion before (Supplementary material).

Because of multicollinearity, we were led to counterintuitive estimates of regression parameters in the CR model including all covariates (Supplementary material): the coefficient of SIE in autumn was estimated at 0.5 (SE = 0.24) and that of SIE in winter was estimated at -0.5 (SE = 0.21) while these two covariates were significantly positively correlated (r = 0.67, p < 0.01).

When we applied the P2CR approach, the algorithm selected two PCs, namely PC3 ($F_{1,27}$ = 7.34, p = 0.01) at step 1 and PC4 ($F_{1,26}$ = 4.63, p = 0.04) at step 2 (Supplementary material), but never did we pick PC1 as we would have done using a classical approach (Grosbois et al. 2008). PC3 was positively correlated to SIE in summer and negatively correlated to temperature in winter, while PC4 was positively correlated to temperature in spring and negatively correlated to SIE in summer (Supplementary material). Survival increased with increasing values of PC3 (Figure 1), with high values of SIE in summer and low values of temperature in winter (resp. low values of SIE in summer and high values of temperature in winter) corresponding to high (resp. low) survival.

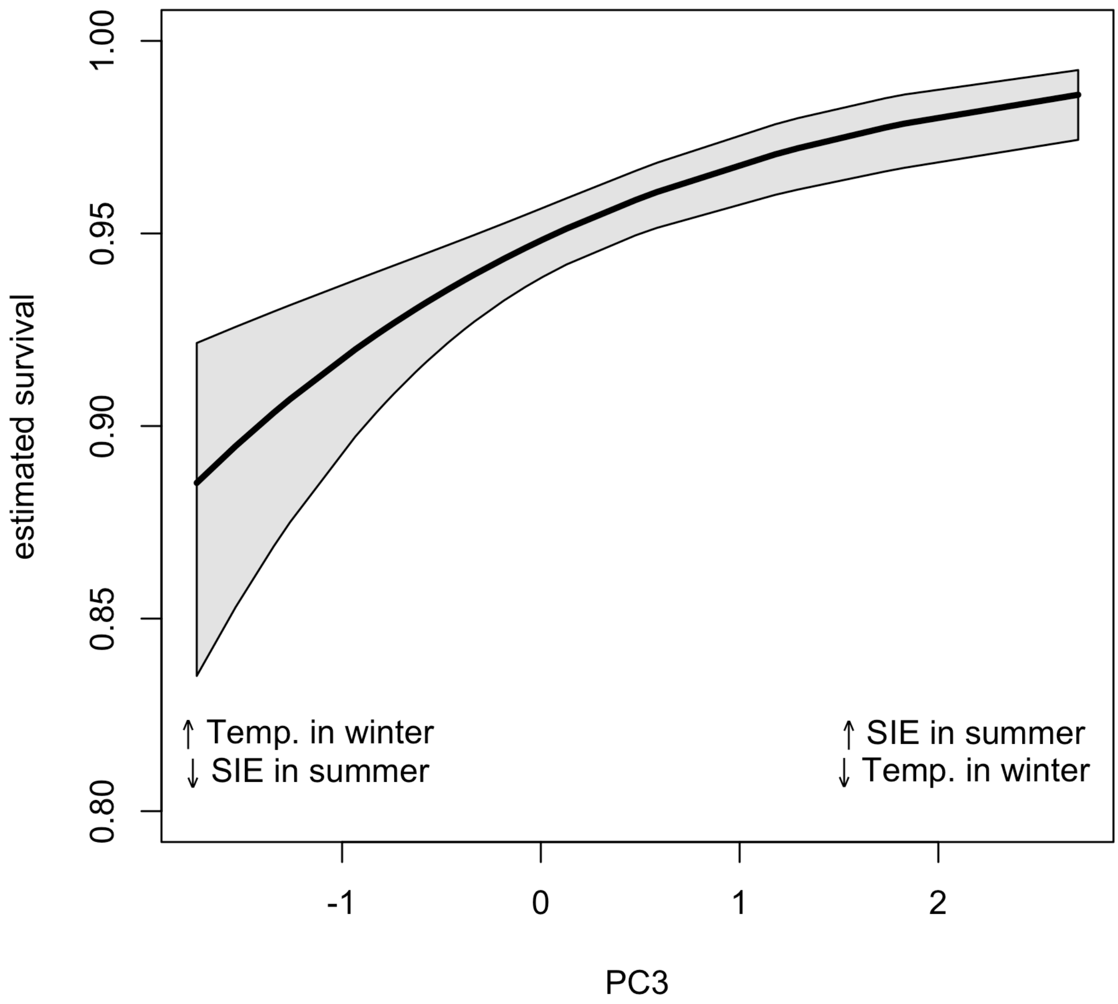

Figure 1: Survival of Snow petrel as a function of PC3.

Survival decreased with increasing values of PC4 (Figure 2), with high values of temperature in spring and low values of SIE in summer (resp. low values of temperature in spring and high values of SIE in summer) corresponding to low (resp. high) survival.

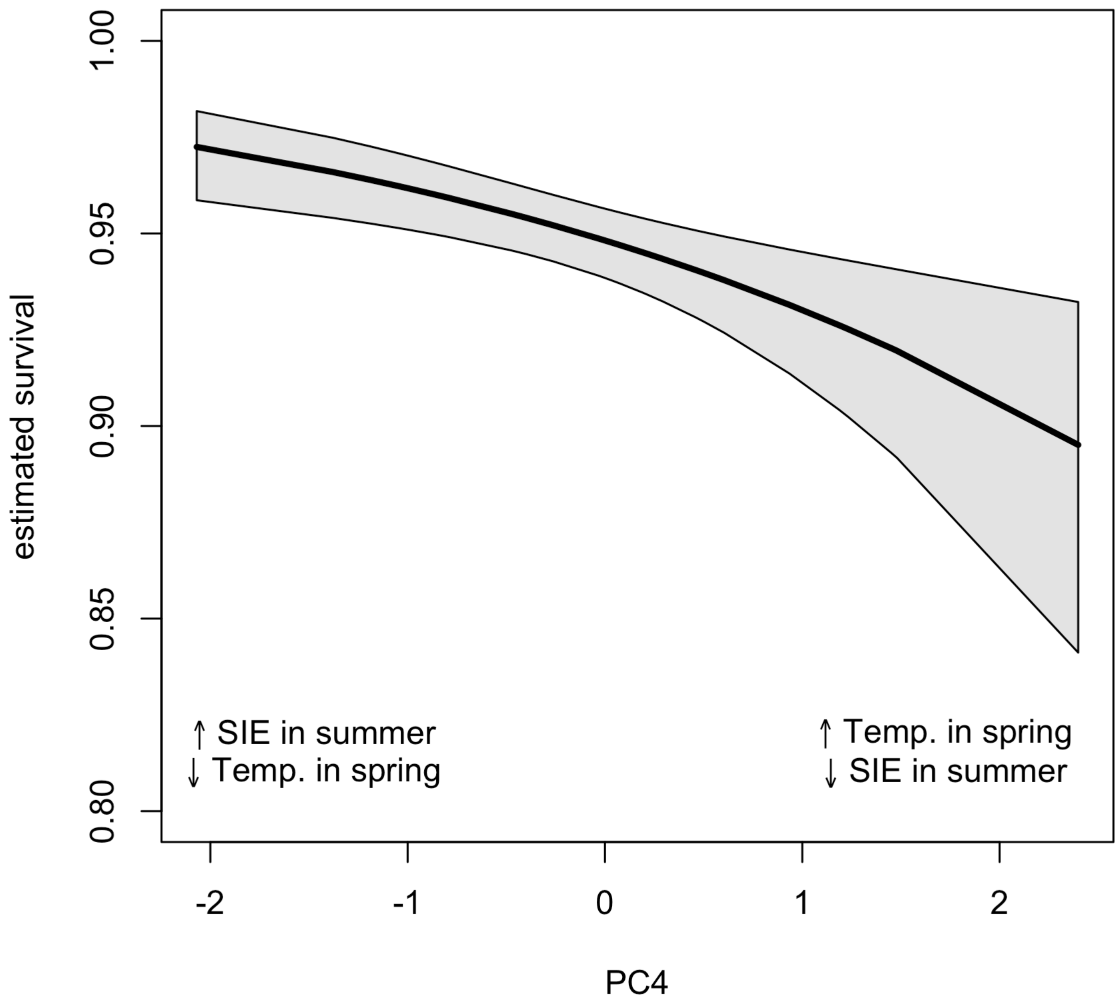

Figure 2: Survival of Snow petrel as a function of PC4.

The P2CR approach also led to more precise survival estimates when compared to the model incorporating all original covariates (Figure 3).

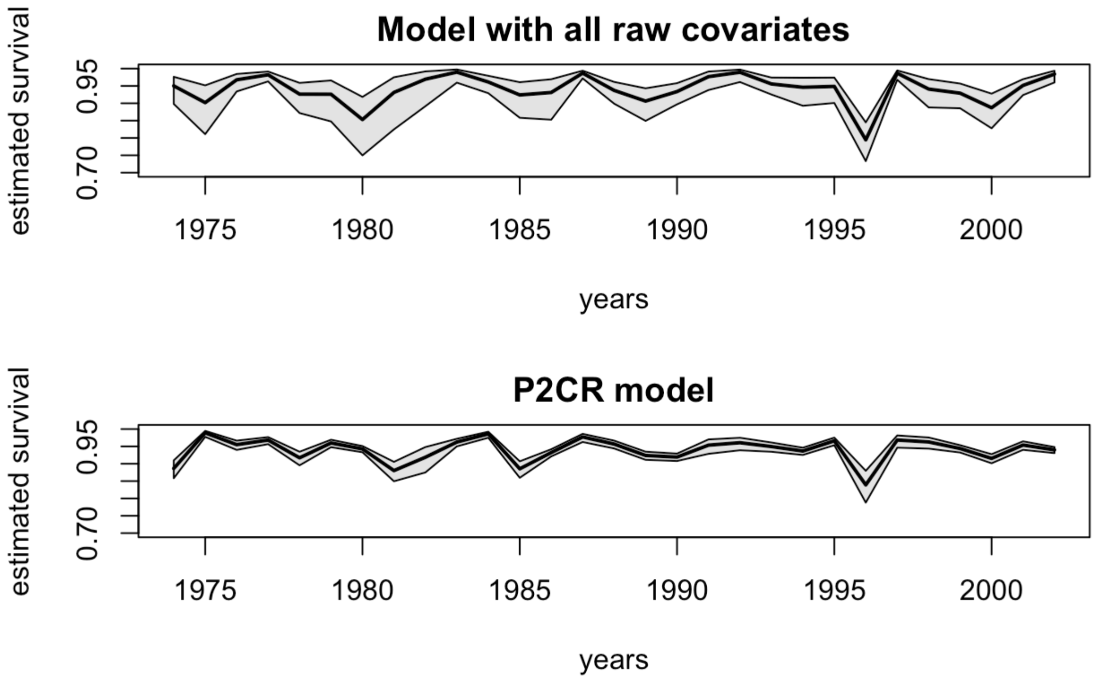

Figure 3: Survival of Snow petrel over time as estimated from the model with all original covariates (top panel) vs. the PC2R model (bottom panel).

Discussion

We introduce a new approach combining principal component analysis and capture-recapture models to deal with many possibly correlated explanatory covariates. Our approach requires standard statistical tools, which allows an efficient and easy implementation using standard software.

Snow petrels and climatic conditions

In summer, snow petrels exclusively forage within the pack-ice tens to hundreds of kilometers from the colony where they catch sea ice-associated species, such as Antarctic silverfish (*Pleuragramma antarcticum*) and Euphausiids, to feed their chick (Ridoux & Offredo 1989). This is an energetically demanding period for breeding adults and, during years with reduced sea-ice extent, food resources may be less abundant and snow petrels may be forced to cover larger distances to find suitable foraging habitats, with potential survival costs. Assuming air temperature was a proxy of sea surface temperature variations, the negative effect of warmer temperatures on survival is coherent with general patterns found between sea surface temperature and demographic parameters in seabirds (Barbraud *et al.* 2012). In many marine ecosystems warmer temperatures are associated with decreased primary production and food resources for top predators. Although the low survival in 1996 corresponded to a year with reduced sea-ice extent in summer, the drop in survival was high and remains unexplained at the moment.

Principal component CR models

When multiple covariates have to be considered to estimate survival, both issues of dimensionality and multicollinearity can lead to biased estimates, inflated precision as well as lack of statistical power. In such a context, the P2CR modeling framework has proved particularly useful in our example, mainly because few PCs were selected which were easily interpretable. We acknowledge that PCs with little interpretability might have been picked up by our method. To make the interpretation easier, PCA results can be post-processed by rotating axes to improve correlations between raw variables and PCs like in the varimax method (Kaiser 1958). Recent developments in the field of multivariate analyses could also be useful, like methods to handle with missing values in PCA (Dray & Josse 2015).

In statistical ecology, one of our objectives is to try and explain variation in state variables such as abundance, survival and the distribution of species. Dimension-reduction methods are promising to deal with many correlated covariates for the analysis of CR or occupancy data.

ACKNOWLEDGEMENTS

We greatly acknowledge all of the wintering fieldworkers involved in the monitoring programs in Terre Adélie since 1962, and Dominique Besson for the management of the database. The study on petrels was supported by Expéditions Polaires Françaises, Institut Paul Emile Victor (program IPEV 109, resp. H. Weimerskirch) and Terres Australes et Antarctiques Françaises.

LITERATURE CITED

Aguilera, A.M., Escabias, M. & Valderrama, M.J. (2006) Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Computational Statistics and Data Analysis*, **50**, 1905–1924.

Barbraud, C., Rolland, V., Jenouvrier, S., Nevoux, M., Delord, K. & Weimerskirch, H. (2012) Effects of climate change and fisheries bycatch on Southern Ocean seabirds: A review. *Marine Ecology Progress Series*, **454**, 285–307.

Barbraud, C., Weimerskirch, H., Guinet, C. & Jouventin, P. (2000) Effect of sea-ice extent on adult survival of an Antarctic top predator: the snow petrel *Pagodroma nivea*. *Oecologia*, **125**, 483-488

Dray, S. & Josse, J. (2015) Principal component analysis with missing values: a comparative survey of methods. *Plant Ecology*, **216**, 657–667.

Graham, M.H. (2003) Confronting multicollinearity in ecological multiple regression. *Ecology*, **84**, 2809–2815.

Grosbois, V., Gimenez, O., Gaillard, J.M., Pradel, R., Barbraud, C., Clobert, J., Møller, A.P. & Weimerskirch, H. (2008) Assessing the impact of climate variation on survival in vertebrate populations. *Biological Reviews*, **83**, 357–99.

Jolliffe, I.T. (2002) Principal Component Analysis, Second Edition. Springer-Verlag, New

York.

Kaiser, H.F. (1958) The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, **23**, 187–200.

Laake, J.L. (2013) RMark: An R Interface for Analysis of Capture-Recapture Data with MARK. AFSC Processed Rep 2013-01, 25p. Alaska Fish. Sci. Cent., NOAA, Natl. Mar. Fish. Serv., 7600 Sand Point Way NE, Seattle WA 98115.

Lebreton, J.-D., Burnham, K.P., Clobert, J. & Anderson, D.R. (1992) Modeling survival and testing biological hypotheses using marked animals: A unified approach with case studies. *Ecological Monographs*, **62**, 67–118.

Pollock, K.H. (2002) The use of auxiliary variables in capture-recapture modelling: an overview. *Journal of Applied Statistics*, **29**, 85–102.

Ridoux, V. & Offredo, C. (1989) The diets of five summer breeding seabirds in Adélie Land, Antarctica. *Polar Biology*, **9**, 137–145.

Skalski, J.R., Hoff, A. & Smith, S.G. (1993) Testing the significance of individual- and cohort-level covariates in animal survival studies. *Marked Individuals in the Study of Bird Population*, eds J.D. Lebreton & P.M. North, pp. 9–28. Birkäuser Verlag, Basel.

White, G.C. & Burnham, K.P. (1999) Program MARK: survival estimation from populations of marked animals. *Bird Study*, **46**, 120–139.