Model selection procedures for high dimensional genomic data

Allan J. Motyer[1] Sally Galbraith[2] Susan R. Wilson[3]

(Received 31 January 2011; revised 22 June 2011)

Abstract

Many complex diseases are thought to be caused by multiple genetic variants. Recent advances in genotyping technology allowed investigators of a complex disease to obtain data for a massive number of candidate genetic variants. Typically each candidate variant is tested individually for an association with the disease. We approach the problem as one of model selection for high dimensional data. We propose a method whereby penalised maximum likelihood estimation provides a reasonably sized set of variants for inclusion in our model. We then perform stepwise regression on this set of variants to arrive at our model. Penalised maximum likelihood estimation is performed with both the lasso and a more recently developed method known as the hyperlasso, with smoothing parameters chosen by cross-validation. The hyperlasso has a penalty function that favours sparser solutions but with less shrinkage of those variables that are included in the model, when compared to the lasso; however, this comes at extra computational cost. We apply the above method to a large genomic data

http://anziamj.austms.org.au/ojs/index.php/ANZIAMJ/article/view/3970
 Published July 20, 2011. ISSN 1446-8735. (Print two pages per sheet of paper.)

set from a previously published mice obesity study and use resample model averaging to assess model performance.

Contents

1 Introduction

Genome-wide association studies (GWAS) have been very successful in recent years in identifying genetic variants (for example, single-nucleotide polymorphisms known as SNPs) that are related to common complex diseases [13]. In GWAS, typically hundreds of thousands of SNPs across the entire genome are genotyped in a number of individuals in order to determine genomic regions associated with the trait of interest. Typically a single-marker approach is used; however, this will only be appropriate when a single SNP is related to the trait. For many complex diseases it is believed that there are many causative genetic factors, in which case a single-marker approach may not be able to identify joint effects of SNPs. For many GWAS for complex traits the SNPs identified have explained only a very small percentage of phenotypic variation [11, 12]. More sophisticated analyses may reveal further insight and "advances in statistical methodology will be central in these developments" [21].

The simultaneous analysis of all SNPs in GWAS for a complex trait has been considered recently [10, 20, 4, 1].

The high dimensionality of GWAS data (with the number of SNPs far exceeding the number of individuals) poses challenges for traditional model selection methods, such as stepwise selection, that use multiple regression with a variable selection procedure. First, the huge number of SNPs causes computational difficulties. Second, such methods are not well suited to deal with the problem of multicollinearity among SNPs.

We approach the problem of model selection (that is, identification of the SNPs that are associated with the phenotype) by making use of penalized-likelihood methods. Fan and Lv [6] reviewed recent advances in variable selection in high dimensional statistical modeling. Here we use the well-known lasso procedure [16] and the more recently introduced variation known as the hyperlasso [10], which is reported to be better able to select true causal variants while keeping a sparse solution.

Penalized-regression approaches have been attracting increased attention recently in the context of GWAS. However, it is still an open question as to how a penalized-likelihood model selection method, such as the lasso or hyperlasso, might best be incorporated into a GWAS analysis. We present a three step procedure for simultaneous analysis of all SNPs in GWAS. Model selection is performed in the first two steps, and model assessment is carried out in the third step. The first step consists of penalized-likelihood variable selection to identify a set of SNPs for further consideration. Here we use both the lasso and hyperlasso for this first step. The second step is to refine the model by using a traditional variable selection method with the candidate SNPs selected in the first step. Here we use stepwise regression. We check against over-fitting in the first two steps by using cross-validation (based on the model selected by the combination of the first two steps) to determine the smoothing parameters for the first step, which in turn determines the number of SNPs that are selected.

The third step involves assessment of the performance of our model selection

procedure with resample model averaging. This involves taking bootstrap samples of the data, say 100, and, as the name suggests, calculating the proportion of bootstraps for which each SNP is selected when the model selection procedure is applied to the bootstraps. The use of resample model averaging has been shown to improve false discovery rates for high dimensional data [17, 14]. We compare the SNPs obtained by our model selection procedure and the results of resample model averaging to assess model stability—that is, whether the model is likely to change with small changes in the data.

We apply our method to data from a previous study into mice obesity [18], and make comparisons with a simple approach which aims to identify joint SNP effects using the results of a single SNP analysis.

2 Model selection procedure

We simultaneously analyse k SNPs typed in n individuals, with $k \gg n$. We formulate the problem as variable selection in a linear regression analysis that includes a covariate for each SNP and a continuous phenotype. That is, we consider the multiple linear regression model,

$$y_j = \beta_0 + \sum_{i=1}^{k} \beta_i x_{ij} + \epsilon_j \,, \tag{1}$$

where y_j is the phenotypic value of the jth individual, β_0 is a constant, x_{ij} is a variable taking the value 0, 1, or 2 if the genotype of the jth individual at SNP i is homozygous in the major allele, heterozygous, or homozygous in the minor allele, respectively, β_i is a regression coefficient corresponding to the ith SNP, and ϵ_j is the residual error for the jth individual. Residual errors are assumed to be independent and identically distributed following a mean zero Gaussian distribution. Our goal is to identify the causal SNPs for which the corresponding regression coefficient is non-zero.

Since $k \gg n$ one cannot perform standard maximum likelihood estimation. To prevent model overfitting we employ a regularisation method, where we seek to maximise a penalised form of the likelihood which imposes a constraint on the size of the regression coefficients. Both the lasso [16] and the hyperlasso [10] are considered. The lasso method provides an estimate of the regression coefficients,

$$\hat{\boldsymbol{\beta}}^{\ell} = \arg\max_{\boldsymbol{\beta}} \left[L(\boldsymbol{\beta}) - f_{\ell}(\boldsymbol{\beta}; \xi) \right], \tag{2}$$

where L denotes the log-likelihood for the linear regression model, that is,

$$L(\boldsymbol{\beta}) = -\sum_{j=1}^{n} \left(y_j - \beta_0 - \sum_{i=1}^{k} \beta_i x_{ij} \right)^2 + \text{constant}, \tag{3}$$

and f_{ℓ} is a penalty function with a smoothing parameter ξ. When the penalty function is zero we obtain the maximum likelihood estimate. The lasso uses the penalty function

$$f_{\ell}(\boldsymbol{\beta}; \xi) = \xi \sum_{i=1}^{k} |\beta_i|. \tag{4}$$

The hyperlasso estimates of the regression coefficients, $\hat{\boldsymbol{\beta}}^{h}$, are obtained in the same manner; however, the penalty function is

$$f_h(\boldsymbol{\beta}; \lambda, \gamma) = -\sum_{i=1}^{k} \left[\frac{\beta_i^2}{4\gamma^2} + \log D_{(-2\lambda-1)} \left(\frac{|\beta_i|}{\gamma} \right) \right], \tag{5}$$

where λ and γ are smoothing parameters, and $D_{(\alpha)}(\cdot)$ is the parabolic cylinder function [8, 19], with parameter α. An advantage of the lasso and hyperlasso for model selection is that these penalty functions tend to shrink many of the regression coefficients to zero, leaving relatively few non-zero coefficients.

The inspiration for the hyperlasso penalty function can be seen by viewing the problem in a Bayesian framework. The negative of the penalty corresponds to the log-prior density of the regression coefficients. By Bayes' theorem (after

taking logarithms) the objective function in equation (2) is the log-posterior density of the regression coefficients. By finding the maximum we are then finding the mode of the posterior. Under this interpretation both the lasso and hyperlasso assign independent priors with a density sharply peaked at zero to each of the regression coefficients. For the lasso this is the Laplace density. For the hyperlasso it is the normal exponential gamma density (NEG) [9]. The NEG is a generalisation of the Laplace distribution with an additional parameter, and it can be generated by sampling from a Laplace distribution with parameter drawn from a gamma distribution. The parameters λ and γ can be interpreted as shape and scale parameters, respectively. As λ and γ both increase such that $\xi = \sqrt{2\lambda}/\gamma$ remains constant, the NEG converges to the Laplace distribution with parameter ξ. The potential advantage of the hyperlasso over the lasso is that, as λ decreases, its prior density is steeper near zero and flatter elsewhere. This corresponds to a strong prior belief that there are few true causal variants and little prior knowledge of effect sizes, and leads to a sparse solution with less shrinking of non-zero coefficients.

The lasso has a closed form solution for a linear regression model; however, the hyperlasso does not. We make use of the implementation of these procedures described by Hoggart et al. [10], and downloadable from the web [5]. The posterior density is multi-modal and the mode identified depends on the initial values used in the optimization algorithm and the order in which they are updated. The software starts with all coefficients equal to zero, and allows multiple iterations to be performed, each of which permutes the order in which coefficients are updated. We perform 100 iterations as recommended by Hoggart et al. [10].

We analyse our genomic data sets using both the lasso and hyperlasso procedures to obtain a set of candidate SNPs for inclusion in a multiple linear regression model consisting of those SNPs for which we obtain a non-zero regression coefficient estimate. We then restrict the analysis to the set of candidate SNPs and carry out stepwise regression, starting with no variables in the model and allowing variables to be added to and later deleted from the model on the basis of the Bayesian information criterion (BIC). This analysis

is performed using the R software function "step" [15]. The resulting multiple linear regression model includes SNPs as predictor covariates which suggest genomic regions for further investigation.

The values of the lasso and hyperlasso smoothing parameters used in the analysis are determined by ten-fold cross-validation. The data is randomly partitioned into ten groups of individuals. A model is fit to 9/10ths of the data (the training set) and then assessed on the remaining 1/10th of the data (the validation set). This is repeated ten times, corresponding to each of the ten groups being used as the validation set. We vary the smoothing parameters over a wide range of values. For fixed smoothing parameters, for each fold we fit a prediction model by the procedure described above. That is, first obtaining candidate SNPs using either the lasso or hyperlasso then obtaining a linear model by stepwise regression, as described above. For each fold we calculate the mean squared prediction error, and then obtain an estimate of the cross-validation error by averaging over the ten-folds. In the case of the hyperlasso, cross-validation is performed over two smoothing parameters. In each case the smoothing parameters correspond to approximately fifteen SNPs being selected by the lasso and hyperlasso.

2.1 Stability analysis

We perform "resample model averaging" [14, 17] to validate the stability of the model selected using the lasso and hyperlasso. This involves repeatedly resampling the data and performing the lasso and hyperlasso procedure (but not subsequent stepwise regression) on each resample. We perform the form of resample model averaging known as "bagging" based on nonparametric bootstrapping, that is, sampling with replacement. We take 100 resamples, each of size equal to the number of individuals in the analysis. We use smoothing parameters identical to those used in our earlier model selection procedure. We calculate the proportion of resamples for which each SNP is selected, which is termed the "resample model inclusion probability" (RMIP).

A plot of the RMIP for all SNPs versus genome position is used to visually assess stability. It is desirable that the SNPs that were selected in our model are those with the highest RMIP, which would indicate that the model is stable with respect to perturbations in the data.

3 Results

We analyse an experimental data set that was the subject of previous investigations [18, 7]. The experimental details are described therein. The data comprises SNP data for 320 F2 mice (162 female, 158 male) generated by crossing *C57BL/6J apoE null* with *C3H/HeJ apoE null* mice. The SNP data consists of genotypes for 1347 SNPs that show variation between the *C57BL/6J* and *C3H/HeJ* strains. The quantitative trait analysed was abdominal fat mass at 24 weeks of age.

Previous investigations with the current data [18, 7] showed major differences in gene expression levels between sexes among the F2 mice used, and therefore we analyse each sex separately. For each sex we obtain two models—one for each of the lasso and hyperlasso—which we compare.

Several data quality control steps are taken. We exclude SNPs with greater than 5% missing data, minor allele frequency less than 0.01, or a Hardy–Weinberg equilibrium p-value less than 10^{-3} for pooled male and female mice. We further exclude those mice with greater than 2% missing SNP data. This left 138 male and 151 female mice and 1180 SNPs. We perform a final filtering step on the SNPs, following the procedure by Carlson et al. [3] using a Linkage Disequilibrium r^2 threshold of 0.9 as suggested by Balding [2], in which SNPs are grouped into highly correlated bins and only one SNP from each bin is retained. After this step 250 SNPs were retained.

Table 1 summarises the final models obtained using the lasso and hyperlasso. The SNPs selected in the lasso and hyperlasso stage show a lack of commonality in selected SNPs between genders. Further, selected SNPs are from twelve

TABLE 1: Comparison of model fits. Summary of the models obtained using each of the lasso, hyperlasso and top-ranked SNPs based on a single-SNP analysis, followed by stepwise regression.

	Lasso	Hyperlasso	Single-SNP
	Males		
No. of candidate SNPs	15	15	15
No. of covariates	9	11	7
Residual standard error	0.380	0.363	0.453
Residual df	128	126	123
Multiple R-squared	0.577	0.620	0.306
Adjusted R-squared	0.547	0.587	0.278
p-value	$< 2E{-}16$	$< 2E{-}16$	$1.1E{-}08$
	Females		
No. of candidate SNPs	14	15	15
No. of covariates	8	9	6
Residual standard error	0.949	0.932	0.973
Residual df	142	141	133
Multiple R-squared	0.376	0.403	0.348
Adjusted R-squared	0.342	0.365	0.319
p-value	$9.7E{-}12$	$2.1E{-}12$	$1.3E{-}10$

distinct chromosomes in males and eleven distinct chromosomes in females, but only five of these chromosomes are in common. Of the SNPs selected in the final models after the application of stepwise regression, there are no common SNPs between genders for the models based on lasso and there is only one SNP in common between genders for the models based on hyperlasso.

We compare our model selection procedure with a similar approach based on ranking SNPs by their individual association with the phenotype. This approach has two stages, with SNPs first ranked by the p-value of a single-SNP linear model rather than using the lasso or hyperlasso. We retain the top fifteen ranked SNPs based on this single-SNP analysis. Stepwise regression is

then applied to these SNPs. The model obtained with this approach is also summarised in Table 1.

Only six of the top fifteen SNPs for males are selected in stage 1 by either lasso or hyperlasso. This number is ten for females. Of those SNPs selected in the final models after stepwise regression, only three SNPs for males and four SNPs for females are in the respective lasso and hyperlasso final models.

The procedure is designed so that a similar number of SNPs are selected as the "Candidate SNPs" in the first stage. The number of SNPs in the final model is less for the "Single-SNP" approach. This is most likely due to multicollinearity among the SNPs selected in the first stage. The residual standard error, R-squared, adjusted R-squared and p-value of each model shows that both the lasso and hyperlasso approaches outperform the single-SNP approach. The hyperlasso approach also has a more modest advantage over the lasso approach. The majority of SNPs selected with the lasso and hyperlasso approaches were common to both. There were fewer SNPs that were common to all three approaches.

3.1 Validation with resample model averaging

The plot of the resample model inclusion proportion (RMIP) for each SNP against genome position for both the lasso and hyperlasso approaches for males is shown as Figure 1 (a similar plot for females is not shown). This plot has the SNPs selected in the final stage of our model selection procedure marked, as well as those SNPs selected by the single-SNP approach and those identified in the previous study of Wang et al. [18].

The plot shows that the majority of SNPs in the lasso and hyperlasso models have high RMIP, which is desirable. This indicates that the SNPs are highly likely to be retained in the model under perturbation of the data (noting that each resample is a bootstrap). The lasso and hyperlasso do not select all of the SNPs with the highest RMIP, which is not necessarily a problem. The

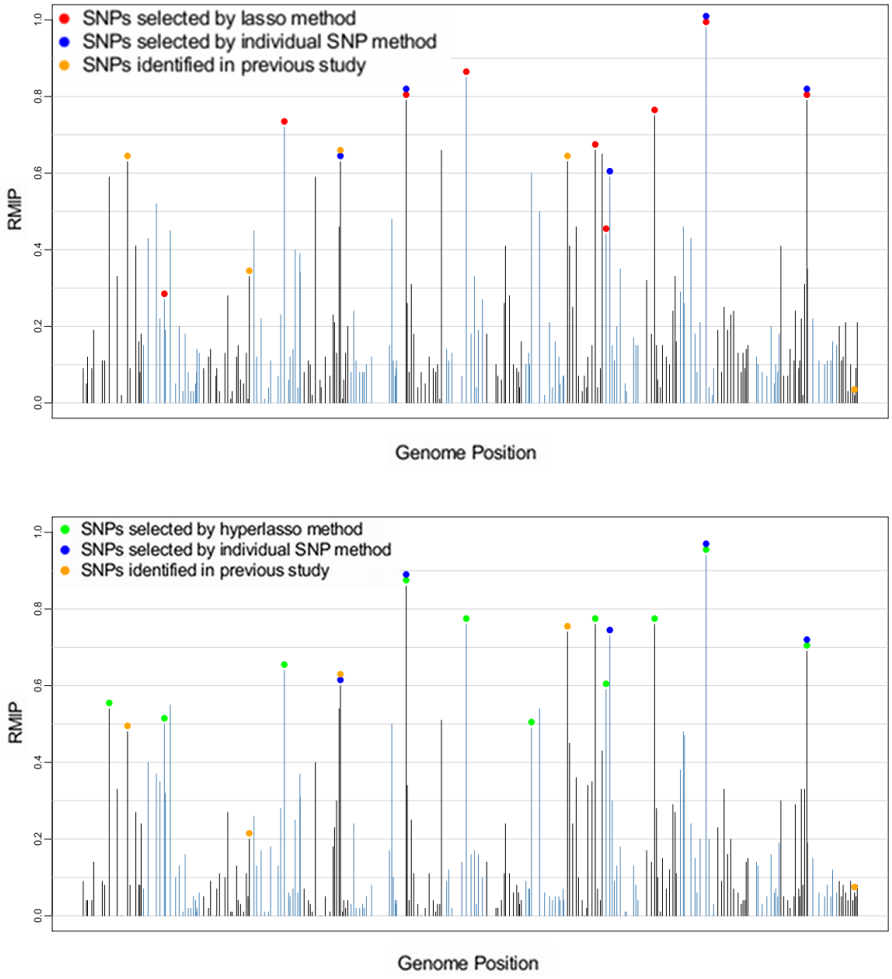

FIGURE 1: The Resample Model Inclusion Proportion (RMIP) for each SNP from bagging (resample model averaging) plotted against SNP genome position, for the male mice. Upper figure is for the lasso approach and lower figure is for the hyperlasso approach. Each chromosome is plotted in alternating colours. The SNPs selected in various models are marked; red and green dots mark SNPs selected by the lasso and hyperlasso methods, respectively; blue dots mark SNPs selected by the individual SNP method; orange dots mark SNPs identified in the previous study [18].

lasso model shows one SNP with RMIP less than 0.4; however, this is not the case for the hyperlasso. This fact, along with the slightly superior model fit displayed in Table 1, indicates that the hyperlasso may be preferable to the lasso.

Acknowledgements We thank Eric Schadt of Pacific Biosciences for providing the data analysed in this article. We also thank Clive Hoggart and David Balding for assistance with queries about the hyperlasso program. We acknowledge funding of this research through Australian National Health and Medical Research Council grant number 525453.

References

[1] Kristin A. Ayers and Heather J. Cordell. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic Epidemiology*, 38:879–891, 2010. doi:10.1002/gepi.20543 C366

[2] David J. Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7:781–791, 2006. doi:10.1038/nrg1916 C371

[3] Christopher S. Carlson, Michael A. Eberle, Mark J. Rieder, Qian Yi, Leonid Kruglyak, and Deborah A. Nickerson. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, 74:106–120, 2004. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1181897/?tool=pubmed C371

[4] Seoae Cho, Kyunga Kim, Young Jin Kim, Jong-Keuk Lee, Yoon Shin Cho, Jong-Young Lee, Bok-Ghee Han, Heebal Kim, Jurg Ott, and Taesung Park. Joint identification of multiple genetic variants via

elastic-net variable selection in a genome-wide association analysis. *Annals of Human Genetics*, 74:416–428, 2010. doi:10.1111/j.1469-1809.2010.00597.x C366

[5] European Bioinformatics Institute. http://www.ebi.ac.uk/projects/BARGEN/. C369

[6] Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20:101–148, 2010. http://www3.stat.sinica.edu.tw/statistica/j20n1/J20N12/J20N12.html C366

[7] Anatole Ghazalpour, Sudheer Doss, Bin Zhang, Susanna Wang, Christopher Plaisier, Ruth Castellanos, Alec Brozell, Eric E. Schadt, Thomas A. Drake, Aldons J. Lusis, and Steve Horvath. Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genetics*, 2:e130, 2006. C371

[8] I. Gradshteyn and I. Ryzik. *Tables of Integrals, Series and Products: Corrected and Enlarged Edition*. Academic Press, New York, 1980. C368

[9] J. E. Griffin and P. J. Brown. Bayesian adaptive lassos with non-convex penalization. Technical report, University of Kent, 2007. http://www2.warwick.ac.uk/fac/sci/statistics/crism/research/working_papers/2007/paper07-2/07-2wv2.pdf C369

[10] Clive J. Hoggart, John C. Whittaker, Maria De Iorio, and David J. Balding. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genetics*, 4:e1000130, 2008. doi:10.1371/journal.pgen.1000130 C366, C368, C369

[11] B. Maher. Personal genomes: The case of the missing heritability. *Nature*, 456:18–21, 2008. doi:10.1038/456018a C365

[12] T. A. Manolio et al. Finding the missing heritability of complex diseases. *Nature*, 461:747–753, 2009. doi:10.1038/nature08494 C365

[13] Mark I. McCarthy, Goncalo R. Abecasis, Lon R. Cardon, David B. Goldstein, Julian Little, John P. A. Ioannidis, and Joel N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9:356–369, 2008. doi:10.1038/nrg2344 C365

[14] Nicolai Meinshausen and Peter Buehlmann. Stability selection. *Journal of the Royal Statistical Society, Series B*, 72:417–473, 2010. doi:10.1111/j.1467-9868.2010.00740.x C367, C370

[15] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0. http://www.r-project.org/ C370

[16] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996. http://www.jstor.org/stable/2346178 C366, C368

[17] William Valdar, Christopher C. Holmes, Richard Mott, and Jonathan Flint. Mapping in structured populations by resample model averaging. *Genetics*, 182:1263–1277, 2009. doi:10.1534/genetics.109.100727 C367, C370

[18] Susanna Wang, Nadir Yehya, Eric E. Schadt, Hui Wang, Thomas A. Drake, and Aldons J. Lusis. Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. *PLoS Genetics*, 2:e15, 2006. doi:10.1371/journal.pgen.0020015 C367, C371, C373, C374

[19] E. T. Whittaker. On the functions associated with the parabolic cylinder in harmonic analysis. *Proc. London Math. Soc.*, 35:417–427, 1902. doi:10.1112/plms/s1-35.1.417 C368

[20] Jian Yang, Beben Benyamin, Brian P. McEvoy, Scott Gordon, Anjali K. Henders, Dale R. Nyholt, et al. Common SNPs explain a large

proportion of the heritability for human height. *Nature Genetics*, 42:565–569, 2010. doi:10.1038/ng.608 C366

[21] Gang Zheng, Jonathan Marchini, and Nancy L. Geller. Introduction to the special issue: Genome-wide association studies. *Statistical Science*, 24:387, 2009. doi:10.1214/09-STS310 C365

Author addresses

1. **Allan J. Motyer**, Prince of Wales Clinical School, University of New South Wales, NSW 2052, Australia.
mailto:a.motyer@unsw.edu.au

2. **Sally Galbraith**, Prince of Wales Clinical School and School of Mathematics and Statistics, University of New South Wales, NSW 2052, Australia.
mailto:sally.galbraith@unsw.edu.au

3. **Susan R. Wilson**, Prince of Wales Clinical School and School of Mathematics and Statistics, University of New South Wales, NSW 2052, Australia.
mailto:sue.wilson@unsw.edu.au