# SAMPLING BEHAVIOUR OF REDUCED RANK ESTIMATES OF GENETIC COVARIANCE FUNCTIONS

**Karin Meyer**

Animal Genetics and Breeding Unit[1], University of New England, Armidale, NSW 2351

## SUMMARY

A simulation study investigating relative errors and sampling variances of reduced rank estimates of genetic covariance functions from random regression analyses estimating the leading principal components only, is presented. The example considered pertains to covariance functions for growth of beef cattle. It is demonstrated that the leading principal components are estimated most accurately, and that reduced rank estimates yield estimates of covariance functions with similar errors than full rank estimates. Furthermore, it is shown that substantial repartitioning between genetic and permanent environmental covariances can occur if either is modelled with too few principal components. Results emphasize the need for a judicious choice among the possible combinations of rank of fit for different covariance functions.

**Keywords :** Reduced rank, sampling variances, bias, random regression, covariance function

## INTRODUCTION

Reduced rank estimation via the leading genetic principal components has been proposed for analyses involving multiple, possibly highly correlated genetic effects (Kirkpatrick and Meyer 2004). This can reduce the number of parameters to be estimated, and yield more accurate estimates. A typical application are random regression (RR) analyses, which model trajectories through sets of correlated RR coefficients specific to each individual. Simulation studies have shown reduced rank estimation to be advantageous for such analyses (James *et al.* 2000; Kirkpatrick and Meyer 2004; Meyer and Kirkpatrick 2005), but have considered a single source of variation only. In practice, we have repeated records per individual and need to estimate both genetic and permanent environmental effects of the animal and their covariance functions. Depending on the numbers of principal components (PC) fitted for each effect, this can lead to a repartitioning of variation. This paper presents a simulation study investigating biasses and sampling variances for reduced rank estimates of two covariance functions (CF) fitted to model growth of beef cattle.

## MATERIAL AND METHODS

Population values for the simulation were estimates of genetic and permanent environmental CFs and temporary environmental variances for growth of beef cattle from birth to 820 days of age, obtained by Meyer (2005). CF estimates were obtained fitting RRs on quadratic B-spline functions of age at recording with knots at 0, 160, 320, 480, 640 and 821 days. This yielded 7 RR coefficients and, at full rank, 28 covariances among RR coefficients for each CF. Temporary environmental variances were considered independently distributed with changes in variance with age modelled by a step function with 18 classes (0, $61-90$, $91-120$, ..., $271-300$, $301-360$, ..., $721-780$ and $781-820$ days). Matrices of crossproducts were sampled from a Wishart distribution assuming a simple balanced half-sib design, consisting of 1000 unrelated sires with 8 progeny each, and 7 records per progeny, i.e.,

[1] AGBU is a joint venture of NSW Department of Primary Industries and the University of New England

56,000 records in total. The first record for each animal was assumed to be taken at the lowest age, corresponding to a birth weight. The remaining 6 records were distributed at equal intervals of 126 days, but staggered evenly within each progeny group, with ages for the second record from 61 to 188 days and ages for the last record from 694 to 820 days for progeny 1 to 8, respectively. All progeny groups were assumed to have the same age structure, resulting in a total of 48 different ages at recording. No fixed effects were simulated. Let $\mathbf{M}$ denote the sample of cross-products, $\mathbf{V}$ the covariance matrix for observations on a progeny group, and $s$ the number of groups. With unrelated families of the same structure, the likelihood is $-2\log\mathscr{L} = const. + s\log|\mathbf{V}| + \mathrm{tr}\left(\mathbf{V}^{-1}\mathbf{M}\right)$ and is readily maximised (Thompson 1976). Estimates of CFs and temporary environmental variances were obtained by maximum likelihood, considering the first $m = 1,\ldots,7$ genetic PCs. Permanent environmental CFs were, in turn, fitted : a) considering $m$ PCs as for the genetic CF ("Equal"), b) considering $m+1 \leq 7$ PCs ("Plus 1"), and c) fitting all 7 PCs ("Full"). Additional simulations were carried out assuming only genetic or permanent environmental variances affected observations. Accuracy of estimates was measured as relative error for eigenvalues $\lambda_i$ and the genetic CF, $\mathscr{G}$ evaluated by numerical integration. Relative errors (RE) in genetic eigenfunctions, $\mathbf{e}_i$, were measured as the angle between 'true' and estimated eigenfunctions; see Kirkpatrick and Meyer (2004).

$$RE(\hat{\lambda}_i) = \frac{\hat{\lambda}_i}{\lambda_i} - 1 \qquad RE(\hat{\mathscr{G}}) = n^{-2}\sum_{r=1}^{n}\sum_{s=1}^{n}\left(|\hat{\mathscr{G}}(r,s)/\mathscr{G}(r,s)|\right) - 1 \qquad RE(\hat{\mathbf{e}}_i) = \frac{180}{\pi}\,arccos\frac{\hat{\mathbf{e}}_i'\mathbf{e}_i}{\|\hat{\mathbf{e}}_i\|\,\|\mathbf{e}_i\|}$$

where '$\hat{\cdot}$' denotes an estimate, $\|.\|$ is the vector norm, $\mathscr{G}(r,s)$ is the genetic covariance between ages $r$ and $s$, and $n = 101$ is the grid size used for numerical evaluation of the genetic CF and its eigenfunctions. A total of 4000 replicates were simulated for each scenario.

## RESULTS AND DISCUSSION

Average (log) likelihood ($\log\mathscr{L}$) values and corresponding Bayesian Information Criteria (BIC) for the different analyses are summarised in Table 1. As commonly found, $\log\mathscr{L}$ favoured more detailed models, increasing significantly until at least 6 genetic PCs were fitted. Involving a stringent penalty for the number of parameters, BIC was lowest for a model fitting 4 genetic PCs, provided the permanent environmental CF was estimated involving at least one more PC than for the genetic CF ("Plus 1" or "Full").

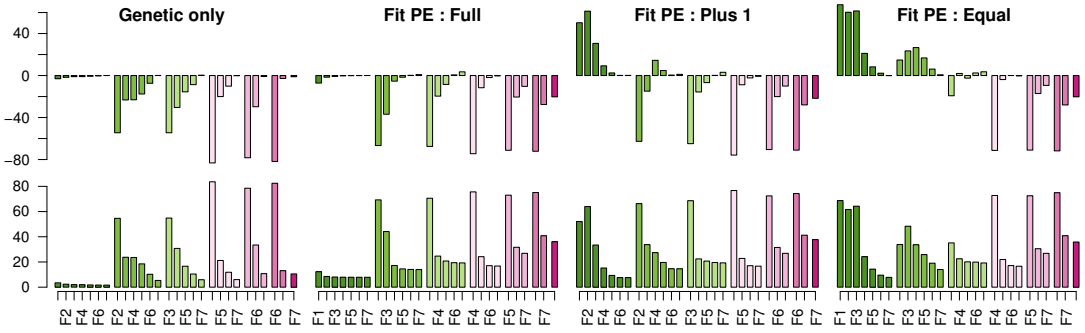


**Figure 1. Relative error in estimates of genetic eigenvalues (top) from analyses fitting $m$ genetic principal components (F$m$), together with root means square errors (bottom); both in %**

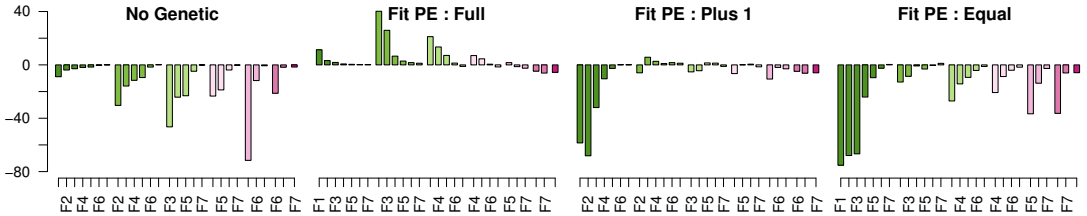


**Figure 2. Relative error in estimates of permanent environmental eigenvalues (in %).**

**Table 1. Log likelihood** $(\log \mathscr{L})^A$ **and corresponding Bayesian Information Criteria (BIC)**

| | "Full" | | | "Plus 1" | | | "Equal" | | |
|---|---|---|---|---|---|---|---|---|---|
| Fit | $p^B$ | $\log \mathscr{L}$ | BIC | p | $\log \mathscr{L}$ | BIC | p | $\log \mathscr{L}$ | BIC |
| 1 | 53 | -188.0 | 955.4 | 38 | -867.0 | 2149.4 | 32 | -1764.3 | 3878.5 |
| 2 | 59 | -103.4 | 851.9 | 49 | -272.6 | 1080.9 | 44 | -507.5 | 1496.0 |
| 3 | 64 | -52.7 | 805.2 | 58 | -93.2 | 820.6 | 54 | -163.3 | 917.1 |
| 4 | 68 | -21.6 | 786.7 | 65 | -33.8 | 778.2 | 62 | -55.5 | 788.8 |
| 5 | 71 | -8.1 | 792.4 | 70 | -10.4 | 786.0 | 68 | -19.7 | 782.9 |
| 6 | 73 | -2.9 | 804.0 | 73 | -2.5 | 803.1 | 72 | -6.3 | 799.7 |
| 7 | 74 | -1.6 | 812.3 | 74 | -1.2 | 811.5 | 74 | -2.3 | 813.6 |

$^A$ +175,200, $^B$ Number of parameters to be estimated

Relative errors in estimates of genetic eigenvalues and square root values of corresponding mean square errors (RMSE) across replicates are shown in Figure 1. Results are grouped according to eigenvalues, i.e., for each of the four analyses there are the 7 estimates of the first eigenvalue, $\lambda_1$, from analyses fitting $1,\ldots,7$ PCs (F1, …, F7), followed by the 6 estimates for $\lambda_2$ from analyses F2 to F7, and so forth till the single estimate of $\lambda_7$ from F7. In the absence of permanent environmental effects or when fitting all PCs ("Full") for the corresponding CF ($\mathscr{P}$), $\lambda_1$ was estimated accurately, regardless of the number of genetic PCs fitted. Subsequent $\lambda_i$ ($i = 2,\ldots,7$) tended to be estimated with large relative errors, especially if they were the last PC fitted (i.e., $\lambda_i$ obtained from F$i$). Population eigenvalues were 944.7, 51.9, 16.1, 9.5, 5.9, 3.2 and 0.2 for $\mathscr{G}$, and 638.3, 85.2, 63.3, 43.8, 14.9, 7.3 and 0.2 for $\mathscr{P}$, i.e., relative importance of PCs tailed off considerably slower for permanent environmental than for genetic effects.

As illustrated in Figure 2, this resulted in substantial downward biasses in estimates of the first eigenvalue of $\mathscr{P}$ if less than 4 permanent environmental PCs were fitted. This caused some of the permanent environmental variance to be partitioned into the genetic components, causing marked upwards biasses, especially in $\lambda_1$ (see Figure 1). Relative errors in estimates of $\lambda_4$ to $\lambda_7$ differed little between analyses, suggesting that errors were not attributable to repartitioning of permanent environmental variation.

In addition, temporary environmental variances (not shown) tended to 'pick up' some of the unex-

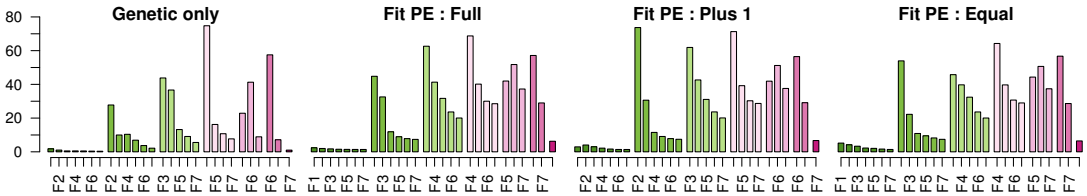


**Figure 3. Error in estimates of genetic eigenfunctions (in degrees).**

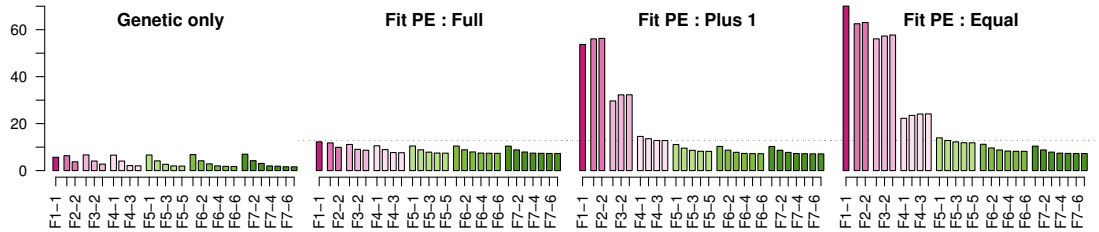


**Figure 4. Relative error in estimates of genetic covariance function (in %)**

plained variation when reduced numbers of PCs were fitted. RMSE errors were of similar magnitude than relative errors, indicating that a large proportion reflected bias rather than sampling variation. Simulating a sizable, well structured data set, no 'intermediate optimum' in RMSE, as might be expected, was apparent, i.e., additional sampling variances due to an increase in the number of parameters to be estimated were small enough as not to outweigh the reduction in bias with increasing number of PCs considered. This may be different for smaller or less well structured data sets (Kirkpatrick and Meyer 2004).

Errors in estimates of genetic eigenfunctions are shown in Figure 3. Whilst estimates of the first genetic eigenvalue were biassed when permanent environmental effects were modelled by too few PCs, the corresponding estimates of the first genetic eigenfunction were little affected, i.e., repartitioning between CFs affected the estimates of the amount of variation rather than the estimates of the direction of the first PC. As for eigenvalues, subsequent eigenfunctions $\mathbf{e}_i$ were estimated with large errors when they pertained to the last PC fitted ($m = i$). At full rank (F7), average deviations in estimates of $\mathbf{e}_i$ tended to increase with $i$, i.e., while the first and second eigenfunction could be estimated accurately, subsequent eigenfunctions tended to be subject to substantial errors. Fortunately, the latter explained little variation, so that estimates of the genetic CF $\mathscr{G}$ were dominated by the first two PCs, and relatively little affected. This is illustrated in Figure 4, which shows relative errors in estimates of $\mathscr{G}$. Here 'F$i-j$ denotes an analysis fitting $i$ genetic PCs, but considering the first $j \leq i$ PCs only in constructing $\hat{\mathscr{G}} = \sum_{r=1}^{j} \hat{\lambda}_r \hat{\mathbf{e}}_r' \hat{\mathbf{e}}_r$. As indicated by the BIC (see Table 1), there were only small reductions in error for analyses fitting more PCs over that for an analysis fitting 4 genetic and 5 permanent environmental PCs.

## CONCLUSIONS

Reduced rank estimation of CFs in RR analyses can yield highly parsimonious models, at negible loss in accuracy. As shown, estimates of CFs are dominated by the leading PCs which are estimated accurately, provided all sources of variation are modelled adequately. Subsequent PCs, contributing little variation, cannot be estimated reliably. The observed repartitioning between sources of variation emphasizes the need to choose the combination of numbers of PCs fitted for all CFs modelled carefully.

## REFERENCES

James, G. M., Hastie, T. J. and Sugar, C. A. (2000) *Biometrika* **87**:587.
Kirkpatrick, M. and Meyer, K. (2004) *Genetics* **168**:2295.
Meyer, K. (2005) *Genet. Select. Evol.* **37**. (in press).
Meyer, K. and Kirkpatrick, M. (2005) *Proc. Roy. Soc. B* **360**:000. (in press).
Thompson, R. (1976) *Biometrics* **32**:903–917.