هفتمین کنگره علوم دامی ایران

صحّت ارزش‌های اصلاحی ژنومی پیش بینی شده با استفاده از تابعیت مؤلفه‌های اصلی به روش بیزی

سید مهدی حسینی وردنجانی [۱]، محمد مهدی شریعتی [۲*]

۱. دانشجوی دکتری ژنتیک و اصلاح نژاد دام دانشگاه فردوسی مشهد

۲. هیئت علمی گروه علوم دامی دانشگاه فردوسی مشهد

ایمیل نویسنده مسئول: shariati52@gmail.com

چکیده

حساسیت مدل‌های آماری پیش بینی ژنومی به انتخاب پرایور مناسب برای پارامترهای کلیدی به خوبی اثبات شده است. در انتخاب ژنومی با استفاده از رگرسیون مؤلفه‌های اصلی (PCR)، واریانس یک کمیّت ثابت تلقی شده و به طور مساوی و یا متناسب با مقادیر ویژه، به هر PC اختصاص داده می‌شود. در این مطالعه اثرات و واریانس PCs با اختصاص توزیع پیشین تصادفی تلقی شد و روش PCR جدیدی ارائه شد. صفتی با وراثت پذیری ۰/۲۵ بر روی ژنومی دارای سه کروموزوم با ۳۰۰۰SNP و ۶۰ QTL شبیه سازی شد. سه توزیع مختلف شامل نرمال، t مقیاس شده و نمایی مضاعف برای PC اسکورها و توزیع کای اسکویر معکوس مقیاس شده برای واریانس‌ها در نظر گرفته شد. همچنین PCR با واریانس متناسب با مقادیر ویژه (PCR-Eigen) و روش‌های بیز ریج (BRR)، بیز A (BA) و بیز B (BB) نیز بررسی شدند. نتایج نشان داد که ضرایب همبستگی پیرسون و رتبه‌ای بین ارزش‌های اصلاحی واقعی و ژنومی پیش بینی شده با استفاده از توزیع‌های پیشین مختلف به ترتیب ۳/۲-۲/۳ و ۲/۹-۲/۷ درصد بالاتر از PCR-Eigen است. صحّت و اریبی پیش بینی‌ها با اختصاص توزیع پیشین به پارامترها مشابه روش‌های BR، BA و BB بود ولی با PCR-Eigen حداقل ۳/۳ درصد پایین‌تر بود. PCR-Eigen به صورت ترتیبی به چند PC اول وزن بیشتری داد و PCs بعدی تقریباً وزنی معادل صفر داشتند در حالیکه با اختصاص توزیع پیشین به پارامترها، به مجموعه‌ی PCهای اول وزن بیشتری داده شد و وزن PCs بعدتر نیز هنوز قابل توجّه بود.

کلمات کلیدی: انتخاب ژنومی- مدل‌های آماری- توزیع پیشین- مقادیر ویژه

مقدمه

انتخاب ژنومی مبتنی بر برآورد ارزش‌های اصلاحی با استفاده از نشانگرهای متراکم بوده [۵] و با توجه به مزیت‌های آن و پیشرفت‌های اخیر در تکنولوژی تعیین ژنوتیپ، می‌تواند فرصت خوبی برای بهبود ژنتیکی سریع‌تر در برنامه‌های اصلاح نژادی فراهم کند. با اینکه روش‌های نسبتاً زیادی برای برآورد اثرات نشانگرها پیشنهاد شده [۲] ولی هنوز مشکلات مهّمی از جمله مشکلات آماری در این زمینه وجود دارد. مهّمترین چالش‌های آماری شامل هم راستایی چندگانه ناشی از عدم تعادل پیوستگی بین نشانگرها که باعث ناپایدار شدن برآوردها می‌شود [۱] و مسئله وجود تعداد زیاد نشانگر در برابر تعداد کم مشاهدات می‌شوند.

آنالیز مؤلفه‌های اصلی (PCA) یکی از انواع روش‌های آنالیز چند متغیره است که مهمترین هدف آن کاهش ابعاد داده‌های همبسته با تبدیل آنها به مجموعه‌ای جدید از متغیرهای متعامد است [۳]. متغیرهای جدید ناهمبسته، مؤلفه‌های اصلی (PC)، به نحوی

رتبه‌بندی می‌شوند که چند PC اول بیشترین تغییرات حاضر در مجموعه داده‌های اولیه را بیان می‌کند و سهم PCهای بعدی در تشریح واریانس بسیار کم بوده و بنابراین با حذف این PCها اطلاعات زیادی از دست نخواهد رفت.

مهمترین مسئله در آنالیز رگرسیون با استفاده از PCها، مشخص کردن سهم هر یک از PC اسکورها در تشریح واریانس کل است. مطالعات اولیه از فرض برابر بودن سهم هر یک از PC اسکورها استفاده کردند [۷]. ولی در مطالعات بعدی با فرض متفاوت بودن واریانس PC اسکورها و استفاده از مقادیر ویژه به عنوان پرایورهای واریانس صحّت پیش بینی‌ها افزایش یافت [٤]. با این حال هر دو این روش‌ها واریانس را به عنوان یک متغیر ثابت از قبل معلوم در نظر می‌گیرند در حالی که از قبل معلوم بودن این واریانس قطعی نیست. از طرفی با حذف برخی از PCها، سهم PCهای باقیمانده از واریانس لزوماً مشابه قبل نیست. بنابراین هدف مطالعه حاضر توسعه روشی برای برآورد اثر هر یک از PC اسکورها و به طور طبیعی واریانس مربوط به آنها در چارچوب روش بیز برای غلبه بر مشکلات ذکر شده بود.

مواد و روش‌ها

صفتی با وراثت پذیری ۰/۲۵ بر روی ژنومی شامل سه کروموزوم، هر یک به طول ۱۰۰ سانتی مورگان شبیه سازی شد. در مجموع تعداد ۳۰۰۰ نشانگر SNP و تعداد ٦۰ QTL چند آللی به صورت تصادفی بر روی این ژنوم توزیع شد. اثرات QTLها از یک توزیع گاما با پارامتر شکل ۰/٤ و پارامتر نرخ ۱/٦ نمونه‌گیری شد. برای ایجاد تعادل جهش-رانش ژنتیکی، ابتدا جمعیتی شامل ٤۰۰ حیوان ماده و ۲۰ حیوان نر برای ۱۰۰ نسل آمیزش تصادفی داده شدند و طی این ۱۰۰ نسل به ۱۰۰۰ حیوان افزایش یافتند. این حالت برای ٤۰۰ نسل دیگر ادامه یافت و تعداد حیوانات نر در نسل آخر به ۷۰ رسانده شد. به طور تصادفی ٤٥٥ حیوان ماده و ۳٥ حیوان نر از نسل ٥۰۰ به عنوان حیوانات پایه گذار (نسل صفر) انتخاب شدند. از نسل صفر به بعد حیوانات نسل بعدی از آمیزش حیوانات دارای بالاترین ارزش اصلاحی در نسل فعلی انتخاب می‌شدند که برای ۹ نسل ادامه داشت. برای هر دو جنس نر و ماده یک ارزش فنوتیپی از مجموع ارزش QTLهای آن فرد که یک انحراف محیطی دارای توزیع نرمال با میانگین صفر و واریانس یک به آن افزوده می‌شد شبیه سازی شد. حیوانات نسل ۸ و ۹ به عنوان جامعه آزمون استفاده شدند. برای ارائه نتایج غیر تصادفی، شبیه سازی ژنوم و جمعیت ۱۰ بار تکرار شد.

مدل‌های آماری بررسی شده شامل دو گروه می‌شدند. گروه اول شامل بیز ریج، بیز A و بیز B که از SNPها به عنوان پیش بینی کننده استفاده می‌کنند و گروه دوم شامل ٤ مدل مختلف، توضیحات آنها در ادامه خواهد آمد، که از تعداد کمتری PC score به عنوان پیش بینی کننده استفاده می‌کنند. برای آنالیز داده‌ها از مدل آماری: $y_i = \mu + sex_k + \sum_{j=1}^{m} \boldsymbol{z}_{ij} b_j + e_j$ استفاده شد که در آن y_i ارزش فنوتیپی فرد iام، μ اثر میانگین، sex_k اثر kامین جنس، $\boldsymbol{z}_{ij}$ ضریب jامین پیش بینی کننده در فرد iام، b_j اثر پیش بینی کننده jام و e_j نیز باقیمانده است. مدل‌های مختلف بررسی شده در در رابطه با ضرایب، اثرات و توزیعی که به اثرات اختصاص داده می‌شد ($\boldsymbol{z}$، bها و توزیع bها) متفاوت بودند. در گروه اول با استفاده SNPها، $\boldsymbol{z}$ شامل یکی از کدهای ۱-، ۱ و یا ۰، به ترتیب برای دو ژنوتیپ هموزیگوس و ژنوتیپ هتروزیگوس، است و b هم اثر جایگزینی مربوط به آن را برآورد می‌کند. در گروه دوم با استفاده از PC scoreها، $\boldsymbol{z}$ شامل کد مربوط به PC اسکور و b هم برآورد اثر آن است. در نماد ماتریسی مدل ذکر شده به شکل $\boldsymbol{y} = \boldsymbol{Xs} + \boldsymbol{Zb} + \boldsymbol{e}$ است.

انتخاب تعداد PC بر مبنای تشریح واریانس به میزان ۰/۹۹۹ انجام شد و در رگرسیون مؤلفه‌های اصلی با استفاده از PCاسکورها به عنوان متغیرهای پیش بینی کننده، چهار مدل مختلف بررسی شدند. در مدل اول (PCR-Eigen) از مقادیر ویژه به عنوان پرایورهای

واریانس استفاده شد [٤]. در مدل دوم (PCR-Normal)، یک توزیع پیشین نرمال $pcscore \sim N(0, \boldsymbol{I}\sigma^2_{pcscore})$ به اثرات اختصاص داده شد. در مدل PCR-scaled-t یک توزیع حاشیه‌ای t مقیاس شده برای PC scoreها با توزیع پیشین $pcscore_j \sim N(0, \sigma^2_{pcscore_j})$ در نظر گرفته شد. در مدل PCR-Lasso نیز به PC اسکورها یک توزیع حاشیه‌ای نمایی مضاعف با توزیع‌های پیشین $p(pcscore_j, \tau_j^2, \lambda^2|\sigma_e^2) = \{\Pi_k N(pcscore_{jk}|0, \tau_{jk}^2 \times \sigma_e^2) Exp\left\{\tau_{jk}^2 \middle| \frac{\lambda^2}{2}\right\}\} \times G(\lambda^2|r,s)$ اختصاص داده شد. به واریانس(های) PCs توزیع $\chi^{-2}(\sigma^2_{pcscore_{(j)}}|S_{pcscore_{(j)}}, df_{pcscore_{(j)}})$ اختصاص داده شد. در تمامی روش‌ها به واریانس باقیمانده توزیع کای اسکویر معکوس مقیاس شده و به اثرات ثابت یک توزیع یکنواخت اختصاص داده شد [٦].

نتایج و بحث

میانگین تعداد SNP بعد از مرحله کنترل کیفیت ۲۸۷۱/۱ و میانگین تعداد PC برای تشریح ۰/۹۹۹ واریانس ۱۵۹۵/۷بود که در حدود ۵۰٪ کمتر از تعداد SNP است. آماره‌های توصیفی همبستگی پیرسون بین ارزش‌های اصلاحی واقعی و ارزش‌های اصلاحی ژنومی پیش بینی شده با مدل‌های مختلف در جدول ۱ آورده شده است. در بین تمام روش‌های مورد مطالعه بالاترین ضریب همبستگی با بیز B بدست آمد که اختلاف آن با بیز A و بیز ریج بسیار کم بود. بالاترین ضریب همبستگی در روش‌های مبتنی بر PC با اختصاص یک توزیع نرمال به اثرات PC اسکورها و یک توزیع کای اسکویر معکوس مقیاس شده به واریانس آنها بدست آمد. روش‌های بیزی مبتنی بر PC به غیر PCR-Lasso در مقایسه با روش‌های بیزی مبتنی بر SNP منجر به برآورد مشابه ضرایب همبستگی شدند با این که این روش‌ها نسبت به روش‌های مبتنی بر SNP در حدود ۵۰٪ متغیر کمتری داشتند. این در حالی است که ضرایب همبستگی با استفاده از روش PCR-Eigen در مقایسه با روش‌های مبتنی بر SNP تقریباً ٤٪ پایین‌تر بود.

جدول ۱- آماره‌های توصیفی همبستگی پیرسون بین ارزش‌های اصلاحی واقعی و ارزش‌های اصلاحی ژنومی پیش بینی شده

	BRR	BA	BB	PCR-Normal	PCR-t	PCR-Lasso	PCR-Eigen
کمینه	۰/۵۰	۰/۵۰	۰/٤۷	۰/۵۱	۰/٤۸	۰/٤٦	۰/٤۹
بیشینه	۰/۷۸	۰/۷۹	۰/۷۹	۰/۷۹	۰/۷۸	۰/۷۸	۰/۷٦
میانگین	۰/٦۸٤±۰/۰۳	۰/٦۸۵±۰/۰۳	۰/٦۸٦±۰/۰۳	۰/٦۸۳±۰/۰۳	۰/٦۸۰±۰/۰۳	۰/٦۷٤±۰/۰۳	۰/٦۵۱±۰/۰۳
واریانس	۰/۰۰۷	۰/۰۰۷	۰/۰۰۸	۰/۰۰۷	۰/۰۰۸	۰/۰۰۹	۰/۰۰۷

BRR: بیز ریج، BA: بیز A، BB: بیز B، PCR-Normal: رگرسیون مؤلفه‌های اصلی با توزیع حاشیه‌ای نرمال، PCR-t: رگرسیون مؤلفه‌های اصلی با توزیع حاشیه‌ای t مقیاس شده، PCR-Lasso: رگرسیون مؤلفه‌های اصلی با توزیع حاشیه‌ای نمایی مضاعف، PCR-Eigen: رگرسیون مؤلفه‌های اصلی با استفاده از مقادیر ویژه به عنوان پرایور واریانس

ضریب رگرسیون ارزش‌های اصلاحی واقعی بر ارزش‌های اصلاحی پیش بینی شده با مدل‌های مختلف به عنوان معیار اریبی پیش بینی‌ها در جدول ۲ نشان داده شده است. برآوردهای نا اریب با استفاده از بیز A با شیب رگرسیون معادل ۱ بدست آمد. مدل‌های بیزی مبتنی بر PC نسبت به بیز A به مقدار ناچیزی اریب بوده ولی هنوز ضرایب رگرسیون آنها در دامنه‌ای قابل قبول با متناظرهای خود با استفاده از SNP هستند. استفاده از PCR-Eigen منجر به برآورد بیش از حد (ضریب رگرسیون= ۰/۷۰٦) ارزش‌های اصلاحی حیوانات شد. در مطالعات قبلی ضرایب رگرسیون با فرض واریانس برابر و واریانس متناسب با مقادیر ویژه برای PC اسکورها به ترتیب ۰/٦۳ و ۰/۸۸ گزارش شدند [٤]. در مطالعه‌ای دیگر با فرض واریانس برابر برای PC اسکورها، ضرایب رگرسیون در تراکم‌های مختلف نشانگر ۰/٦۹۵-۰/٦۵ گزارش شدند [۷].

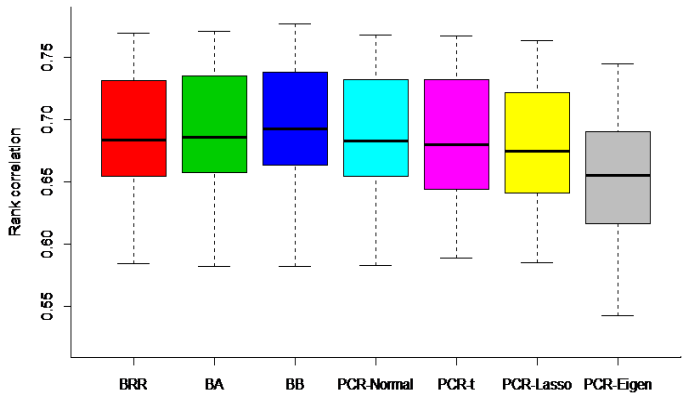
جدول ۲- عرض از مبدأ، ضریب رگرسیون و ضریب تعیین تابعیت ارزش‌های اصلاحی واقعی بر ارزش‌های اصلاحی ژنومی پیش بینی شده

	BRR	BA	BB	PCR-Normal	PCR-t	PCR-Lasso	PCR-Eigen
b0	۰/۸٤±۰/۰۵	۰/۸۳±۰/۰٦	۰/۸٤±۰/۰۵	۰/۷۸±۰/۰۵	۰/۷۹±۰/۰۵	۰/۸۱±۰/۰۵	۰/٦۳±۰/۰۵
b1	۰/۹٦٤±۰/۰۳	۱/۰۰۷±۰/۰٤	۱/۰۹±۰/۰۵	۰/۹٦۲±۰/۰۳	۱/۰۵±۰/۰۵	۱/۰۹±۰/۰۵	۰/۷۰٦±۰/۰۳
R2	۰/٤۷±۰/۰۳	۰/٤۸±۰/۰۳	۰/٤۸±۰/۰٤	۰/٤۷±۰/۰۳	۰/٤۸±۰/۰٤	۰/٤٦±۰/۰۳	۰/٤۳±۰/۰۳

b0: عرض از مبدأ، b1: شیب خط رگرسیون، R2: ضریب تعیین

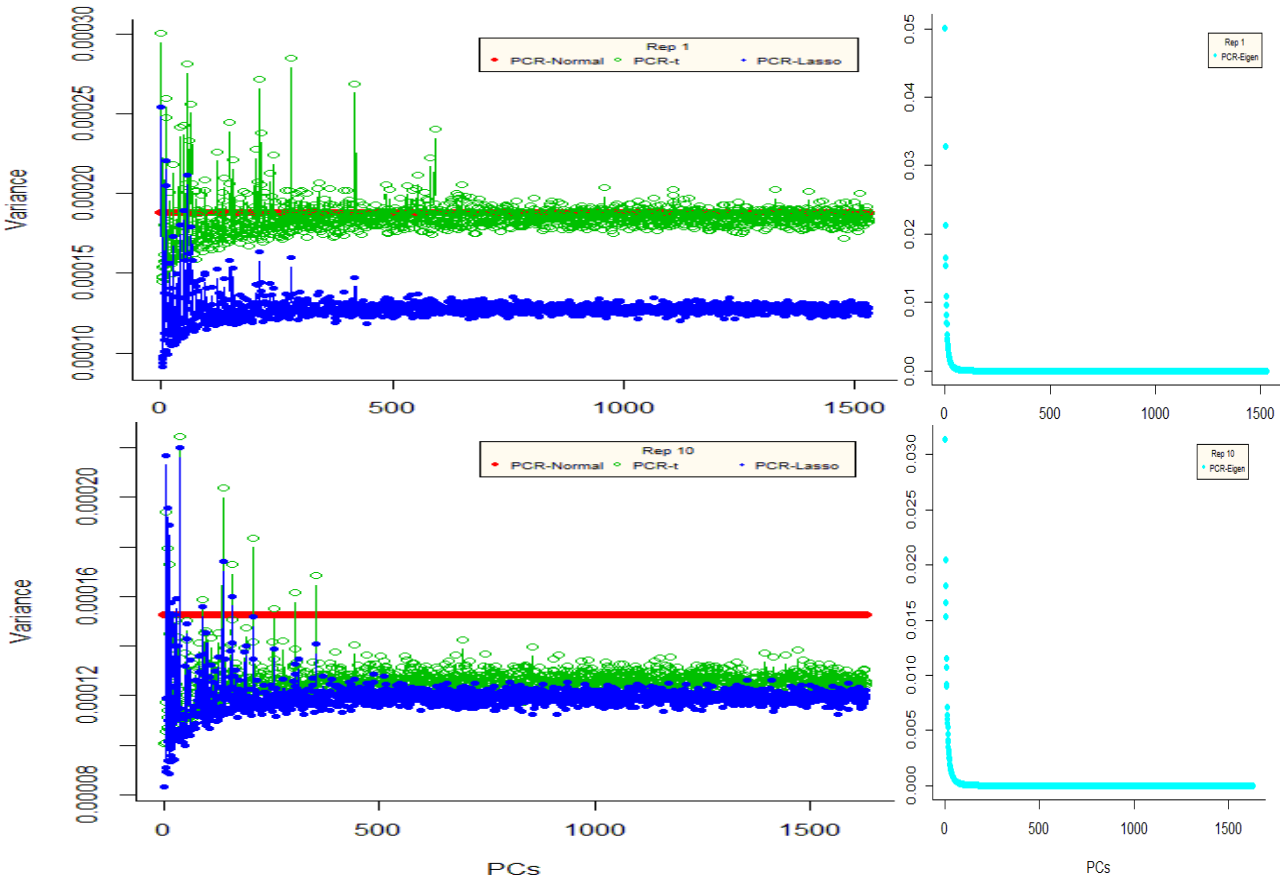
ضریب همبستگی رتبه‌ای بین ارزش‌های اصلاحی واقعی و ارزش‌های اصلاحی ژنومی پیش بینی شده با مدل‌های مختلف در شکل ۱ نمایش داده شده است. روند ضرایب همبستگی رتبه‌ای مشابه موارد متناظر در مورد همبستگی پیرسون است ولی از لحاظ عددی به مقدار قابل توجهی بالاتر هستند. مشابه ضرایب همبستگی پیرسون، تفاوتی بین روش‌های بیزی مبتنی بر PC با روش‌های مبتنی بر SNP وجود ندارد ولی همبستگی بدست آمده با PCR-Eigen پایین‌تر است.

شکل ۱- همبستگی رتبه‌ای بین ارزش‌های اصلاحی واقعی و ارزش‌های اصلاحی ژنومی پیش بینی شده با مدل‌های مختلف



شکل ۲ واریانس PC اسکورها در رگرسیون مؤلفه‌های اصلی مختلف را برای تکرارهای اول و دهم نشان می‌دهد. واریانس نمونه‌گیری شده برای تمام PC اسکورها در مدل PCR-Normal حدود ۰/۰۰۰۲~ بود که اجازه مشارکت مساوی تمام PC اسکورها در تشریح واریانس کل را می‌دهد. مدل PCR-Eigen مطابق با پیش فرض اولیه، بیشترین واریانس (۰/۰۵ و ۰/۰۳۱) را به PC اول و به ترتیب به چند PC اول اختصاص داد و به مؤلفه‌های اصلی بعد از ۲۵۰ واریانس بسیار کوچک نزدیک به صفر اختصاص داد که تقریباً معادل با عدم شرکت آنها در تشریح واریانس کل است. در دو مدل دیگر با واریانس‌های اختصاصی برای هر PC اسکور، تمرکز اصلی به جای PCهای اول بر روی مجموعه‌ای از PCهای اول بود که امکان رتبه بندی مجدد PC به روشی متفاوت از مقادیر ویژه را می‌دهد. از طرفی سهم PCهای بعدی در تشریح واریانس با اختصاص واریانسی بزرگتر از آنچه در PCR-Eigen دارند وجود داشت.

شکل ۲- واریانس اثر PC scoreها در رگرسیون مؤلفه‌های اصلی مختلف در تکرار اول و دهم



منابع:

1- Dadousis, C., Veerkamp, R.F., Heringstad, B., Pszczola, M., Calus, M.P., 2014. A comparison of principal component regression and genomic REML for genomic prediction across populations. Genetics Selection Evolution 46, 1.
2- Garrick, D., Dekkers, J., Fernando, R., 2014. The evolution of methodologies for genomic prediction. Livestock Science 166, 10-18.
3- Jolliffe, I., 2002. Principal component analysis. Wiley Online Library.
4- Macciotta, N.P.P., Gaspa, G., Steri, R., Nicolazzi, E.L., Dimauro, C., Pieramati, C., Cappio-Borlino, A., 2010. Using eigenvalues as variance priors in the prediction of genomic breeding values by principal component analysis. Journal of dairy science 93, 2765-2774.
5- Meuwissen, T., Hayes, B., Goddard, M., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819-1829.
6- Sorensen, D., Gianola, D., 2007. Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer Science & Business Media.
7- Solberg, T.R., Sonesson, A.K., Woolliams, J.A., Meuwissen, T.H., 2009. Reducing dimensionality for prediction of genome-wide breeding values. Genetics Selection Evolution 41, 1.

Accuracy of predicted genomic breeding values using principal component regression with Bayesian method

The sensitivity of statistical genomic prediction models with respect to the choice of the appropriate prior of key parameters is well established. In genomic selection with principal component regression (PCR), variance is treated as a fixed quantity and assigned to PCs equally or proportional to eigenvalues. In this study, effects and variances of PCs treated as random variables by assigning a prior distribution and a new PCR approach was provided. A trait with 0.25 heritability on a genome constructed of 3 chromosome with 60 QTL and 3000 single nucleotide polymorphism was simulated. Three different distribution including normal, scaled-t and double exponential were considered to PC scores and, scaled-inverse chi-square density to variances. PCR with variance proportional to eigenvalues (PCR-Eigen) and, Bayesian Ridge regression (BRR), BayesA (BA) and BayesB (BB) method were also investigated. The results showed that Pearson and Rank correlation coefficients between the true breeding values and estimated genomic breeding values using different prior distribution were 2.3-3.2 and 2.7-2.9 higher than PCR-Eigen, respectively. Prediction accuracy were similar to BRR, BA and BB with assigning prior distribution to parameters, but at least 3.3 smaller by PCR-Eigen. PCR-Eigen gives more weight to first few PCs sequentially and roughly zero weight to later PCs while assigning prior distribution to parameters leads to more weight of first PCs sets and considerable weights to later PCs.

Key words: genomic selection- statistical models- prior distribution- eigenvalues