bayespca Package

Davide Vidotto d.vidotto@uvt.nl

2020-09-09

Contents

bayespca: A package for Variational Bayes PCA

Theoretical background

Principal Components Analysis (PCA) allows performing dimensionality reduction via matrix factorization. While there are several ways to express a PCA model, in what follows will we consider the formulation

$$X = XWP^T + E,$$

where X is a $I \times J$ data matrix (I is the number of units; J the number of continuous variables); W is a $J \times D$ weight matrix ($D \leq J$ is the rank of the reduced matrix); P is the orthogonal loading matrix, such that $P^T P = I_{D \times D}$; and E is an $I \times J$ error matrix. The D principal components can be retrieved with $Z = XW$. In this context, the focus of the inference is typically on W. In particular, when J is large and the main inferential goal is components' interpretation, it is important for the analyst to obtain simple and interpretable components.

The `bayespca` package allows performing the following operations:

1. estimation of the PCA model, with a Variational Bayes algorithm;

2. regularization of the elements of W by means of its prior variances;
3. variable selection, via automatic relevance determination (ARD).

The Variational Bayes algorithm sees the columns of W as latent variables, and P as a fixed parameter. Furthermore, the residuals E are assumed to be distributed according to a Normal distribution with mean 0 and variance σ^2. The following prior is assumed for the d-th column of W:

$$w_d \sim MVN(0, T_d^{-1})$$

where $MVN()$ denotes the density of the Multivariate Normal Matrix, and T_d denotes the prior (diagonal) precision matrix of the d-th component. The j-th element of the diagonal of T_d will be denoted τ_{dj}.

The `bayespca` package

Variational Bayes PCA is implemented through the `vbpca` function, which takes the following arguments as inputs:

- `X` the input matrix;
- `D` the number of components to be estimated;

- `maxIter` the maximum number of iterations for the Variational Bayes algorithm;
- `tolerance` convergence criterion of the algorithm (relative difference between ELBO values);
- `verbose` logical parameter which prints estimation information on screen when `TRUE`;
- `tau` value of the prior precisions; starting value when `updatetau=TRUE` or `priorvar!='fixed'`
- `updatetau` logical parameter denoting whether the prior variances should be updated when `priorvar='fixed'`;
- `priorvar` character argument denoting whether the prior variances should be `'fixed'`, or random with `'jeffrey'` or `'invgamma'` priors;

- `SVS` logical argument which activates Stochastic Variable Selection when set to `TRUE`;
- `priorInclusion` prior inclusion probabilities for the elements of W in the model;
- `global.var` logical parameter which activates component-specific prior variances when set to `TRUE`;
- `control` other control parameters, such as Inverse Gamma hyperparameters (see `?vbpca_control` for more information).

`vbpca` returns a vbpca object, which is a list containing various aspect of the model results. See `?vbpca` for further information. Internally, `vbpca` calls a C++ function (written with Rcpp) to estimate the model. When `nstart`>1, the algorithm will autmatically pick (and output) the best run in terms of final ELBO value.

In what follows, the various estimation modalities allowed by `vbpca` will be introduced. For presentation purposes, a synthetic data matrix with $I = 100$ rows and $J = 20$ columns genereted from three components will be used:

```
set.seed(141)
I <- 100
J <- 20
V1 <- rnorm(I, 0, 50)
V2 <- rnorm(I, 0, 30)
V3 <- rnorm(I, 0, 10)
X <- matrix(c(rep(V1, 7), rep(V2, 7), rep(V3, 6)), I, J)
X <- X + matrix(rnorm(I * J, 0, 1), I, J)
```

I will now proceed with the estimation of the PCA model.

Levels of regularization on the W matrix

Fixed `tau`

With fixed tau, it is possible to specify the model as follows:

```
# Install and load package
# devtools::install_github("davidevdt/bayespca")
library(bayespca)

# De-activate data center and scaling;
ctrl <- vbpca_control(center = FALSE, scalecorrection = -1,
                      plot.lowerbound = FALSE)
# Estimate vbpca with fixed prior variances (equal to 1)
# for the elements of W
mod1 <- vbpca(X, D = 3, maxIter = 1e+03, priorvar = 'fixed',
              control = ctrl, verbose = FALSE )
```

```
## Warning: unscaled data - ELBO values might be positive.
```

```
# Test the class of mod1:
is.vbpca(mod1)
```

```
## [1] TRUE
```

The estimate posterior means of the W matrix can be viewed with:

```r
mod1$muW
```

```
##              Component 1 Component 2   Component 3
## variable 1  -0.376589698 -0.04416511  0.0003399127
## variable 2  -0.373939776 -0.04582346 -0.0111489577
## variable 3  -0.375148656 -0.04305857 -0.0078831833
## variable 4  -0.374770076 -0.04473100 -0.0031124936
## variable 5  -0.376808025 -0.04285791 -0.0100250666
## variable 6  -0.375114064 -0.04446329 -0.0015012122
## variable 7  -0.375069345 -0.04364081 -0.0007181137
## variable 8   0.043916073 -0.37610684 -0.0194987814
## variable 9   0.044338996 -0.37382689 -0.0224165501
## variable 10  0.043216238 -0.37319455 -0.0161965551
## variable 11  0.043432789 -0.37311089 -0.0246530479
## variable 12  0.045420158 -0.37574266 -0.0200072027
## variable 13  0.045158091 -0.37616395 -0.0206149535
## variable 14  0.044605650 -0.37571347 -0.0144837510
## variable 15  0.002905219  0.02229238 -0.4057459196
## variable 16  0.003409761  0.02199152 -0.4068881251
## variable 17  0.003232844  0.02063894 -0.4106993259
## variable 18  0.002919709  0.02319335 -0.4056784549
## variable 19  0.002019259  0.02192116 -0.4088023613
## variable 20  0.001874207  0.02043128 -0.4078307380
```

and the P matrix:

```r
mod1$P
```

```
##              Component 1 Component 2   Component 3
## variable 1  -0.376589904 -0.04416517  0.0003399179
## variable 2  -0.373939981 -0.04582353 -0.0111491289
## variable 3  -0.375148862 -0.04305863 -0.0078833043
## variable 4  -0.374770282 -0.04473106 -0.0031125414
## variable 5  -0.376808232 -0.04285797 -0.0100252205
## variable 6  -0.375114270 -0.04446335 -0.0015012352
## variable 7  -0.375069551 -0.04364087 -0.0007181247
## variable 8   0.043916097 -0.37610735 -0.0194990808
## variable 9   0.044339020 -0.37382740 -0.0224168943
## variable 10  0.043216262 -0.37319506 -0.0161968038
## variable 11  0.043432813 -0.37311139 -0.0246534264
## variable 12  0.045420183 -0.37574317 -0.0200075099
## variable 13  0.045158115 -0.37616446 -0.0206152700
## variable 14  0.044605675 -0.37571398 -0.0144839734
## variable 15  0.002905220  0.02229241 -0.4057521497
## variable 16  0.003409762  0.02199155 -0.4068943728
## variable 17  0.003232846  0.02063897 -0.4107056321
## variable 18  0.002919710  0.02319338 -0.4056846840
## variable 19  0.002019260  0.02192119 -0.4088086384
## variable 20  0.001874208  0.02043131 -0.4078370001
```

Among other things, the function returns the model evidence lower bound (ELBO) and the estimation time:

```r
mod1$elbo
```

```
## [1] -2834.329
```

```r
mod1$time
```

```
##    user  system elapsed 
##       0       0       0
```

Fixed, updatable tau

The prior precisions τ_{dj} can also be updated via Type-II Maximum Likelihood (empirical Bayes updates):

```r
mod2 <- vbpca(X, D = 3, maxIter = 1e+03, priorvar = 'fixed',
                updatetau = TRUE, control = ctrl, verbose = FALSE )
```

```
## Warning: unscaled data - ELBO values might be positive.
```

```r
mod2$muW
```

```
##                 Component 1  Component 2  Component 3
## variable 1  -3.770417e-01 -0.051573499 -0.001788633
## variable 2  -3.747997e-01 -0.025120541 -0.002279085
## variable 3  -3.749896e-01 -0.039169668 -0.002284711
## variable 4  -3.695986e-01 -0.062446420 -0.002188222
## variable 5  -3.798267e-01 -0.031100065 -0.002172447
## variable 6  -3.809607e-01 -0.058646547 -0.001863930
## variable 7  -3.707615e-01 -0.037076450 -0.001858450
## variable 8   4.610643e-02 -0.388243166 -0.019761887
## variable 9   4.122547e-02 -0.376699340 -0.023207406
## variable 10  2.542203e-02 -0.376987920 -0.006935656
## variable 11  5.233439e-02 -0.374340987 -0.022206305
## variable 12  4.536997e-02 -0.374395905 -0.013879307
## variable 13  6.378896e-02 -0.370090650 -0.031591876
## variable 14  3.106493e-02 -0.364598986 -0.002519184
## variable 15  5.847012e-06  0.006198051 -0.406003865
## variable 16  6.041405e-06  0.018464891 -0.407039992
## variable 17  5.235991e-06  0.009013294 -0.411151002
## variable 18  3.518425e-06  0.036121023 -0.398782724
## variable 19  6.288835e-06  0.005824170 -0.410563013
## variable 20  6.214579e-06  0.034945874 -0.412808540
```

The matrix of the prior precisions can be called with

```r
mod2$Tau
```

```
##              Component 1 Component 2  Component 3
## variable 1  6.723601e+00  184.301093 31906.370432
## variable 2  6.797876e+00  437.595910 25321.510595
## variable 3  6.720477e+00  245.899700 25234.633844
## variable 4  6.909797e+00  136.385519 26302.854252
## variable 5  6.614756e+00  336.071863 26671.584537
## variable 6  6.538464e+00  150.961175 30414.866204
## variable 7  6.920149e+00  278.098092 30606.846242
## variable 8  1.973803e+02    6.278101   670.761526
## variable 9  2.288287e+02    6.657350   569.759379
## variable 10 4.149292e+02    6.642324  1989.950540
## variable 11 1.699204e+02    6.739745   597.851184
```

```
## variable 12 2.026152e+02    6.706750   970.392618
## variable 13 1.292935e+02    6.822902   390.944657
## variable 14 3.536290e+02    7.210202  5658.371090
## variable 15 3.158381e+05 2396.464631     5.835080
## variable 16 3.162767e+05  757.091273     5.769939
## variable 17 3.264009e+05 1565.810040     5.674352
## variable 18 3.196566e+05  348.615143     5.995065
## variable 19 3.198439e+05 2526.927990     5.707424
## variable 20 3.182544e+05  363.292432     5.588104
```

Random tau: Gamma prior

It is possible to specify a gamma prior on $\tau_{d,j}$:

$$\tau_{d,j} \sim G(\alpha, \beta)$$

with α shape parameter and β scale parameter. The following code implements an IG(2, .5) prior on the precisions:

```r
# Estimate the model
mod3 <- vbpca(X, D = 3, maxIter = 1e+03,
                priorvar = 'jeffrey', control = ctrl, verbose = FALSE )
```

```
## Warning: unscaled data - ELBO values might be positive.

## Warning: vbpca has not converged. Please re-run by increasing <maxIter> or
## the convergence criterion <tolerance>.
```

```r
mod3$muW
```

```
##               Component 1   Component 2   Component 3
## variable 1  -3.746789e-01 -3.315223e-02 -3.534694e-09
## variable 2  -3.768326e-01 -3.536543e-02 -1.181706e-08
## variable 3  -3.768089e-01 -4.255859e-02 -8.316711e-09
## variable 4  -3.695668e-01 -4.770493e-02 -6.619172e-09
## variable 5  -3.831106e-01 -2.811557e-02 -1.083801e-08
## variable 6  -3.804712e-01 -3.519811e-02 -5.851823e-09
## variable 7  -3.716085e-01 -3.594493e-02 -4.633371e-09
## variable 8   4.224035e-02 -3.791350e-01 -6.406588e-09
## variable 9   3.583886e-02 -3.755145e-01 -1.230848e-08
## variable 10  2.589180e-02 -3.752069e-01  1.958923e-09
## variable 11  3.789386e-02 -3.760433e-01 -6.036710e-09
## variable 12  3.927835e-02 -3.760372e-01 -4.825283e-09
## variable 13  5.345808e-02 -3.776838e-01 -2.914355e-09
## variable 14  2.322450e-02 -3.734752e-01  1.788685e-08
## variable 15 -2.156599e-07 -3.678851e-08 -4.063632e-01
## variable 16 -1.926783e-07 -3.952564e-08 -4.088012e-01
## variable 17 -2.212218e-07 -4.458842e-09 -4.085228e-01
## variable 18 -2.969943e-07  2.194196e-08 -4.083154e-01
## variable 19 -2.114259e-07 -5.940961e-08 -4.093510e-01
## variable 20 -2.070387e-07 -4.261073e-08 -4.081835e-01
```

```r
mod3$Tau
```

```
##              Component 1  Component 2  Component 3
## variable 1  6.809220e+00 3.275482e+02 2.077843e+08
## variable 2  6.730510e+00 3.036210e+02 2.050578e+08
```

```
## variable 3  6.660826e+00 2.300127e+02 2.061375e+08
## variable 4  6.915632e+00 1.998200e+02 2.058102e+08
## variable 5  6.510765e+00 3.927444e+02 2.079856e+08
## variable 6  6.556520e+00 2.948528e+02 2.060881e+08
## variable 7  6.891839e+00 2.979022e+02 2.061982e+08
## variable 8  2.253672e+02 6.572865e+00 5.934440e+07
## variable 9  2.789532e+02 6.702594e+00 5.886445e+07
## variable 10 4.130201e+02 6.708845e+00 5.829204e+07
## variable 11 2.622237e+02 6.697414e+00 5.840853e+07
## variable 12 2.489328e+02 6.657525e+00 5.934046e+07
## variable 13 1.672167e+02 6.575873e+00 5.948928e+07
## variable 14 5.067458e+02 6.893591e+00 5.900699e+07
## variable 15 8.466188e+06 6.290964e+06 5.827806e+00
## variable 16 8.523121e+06 6.260477e+06 5.729407e+00
## variable 17 8.749809e+06 6.438152e+06 5.753097e+00
## variable 18 8.496264e+06 6.208003e+06 5.738715e+00
## variable 19 8.576311e+06 6.397079e+06 5.743597e+00
## variable 20 8.576987e+06 6.326102e+06 5.721094e+00
```

`alphatau` and `betatau` can also be specified as D-dimensional array, in which case the Gamma will have component-specific hyperparameters:

$$\tau_{d,j} \sim G(\alpha_d, \beta_d)$$

.

```
# Set hyperparameter values
ctrl2 <- vbpca_control(center = FALSE, scalecorrection = -1,
                       plot.lowerbound = FALSE,
                       alphatau = 2, betatau = .5)




# Estimate the model
mod4 <- vbpca(X, D = 3, maxIter = 1e+03, priorvar = 'invgamma',
              control = ctrl2, verbose = FALSE )
```

```
## Warning: unscaled data - ELBO values might be positive.
```

```
mod4$muW
```

```
##               Component 1 Component 2   Component 3
## variable 1  -0.376590647 -0.04416831  0.0002942783
## variable 2  -0.373916145 -0.04580796 -0.0111136115
## variable 3  -0.375152371 -0.04306335 -0.0078482119
## variable 4  -0.374771589 -0.04473679 -0.0031289018
## variable 5  -0.376819230 -0.04285286 -0.0099901143
## variable 6  -0.375128398 -0.04445794 -0.0015202569
## variable 7  -0.375056419 -0.04365149 -0.0007404096
## variable 8   0.043918311 -0.37612413 -0.0195192789
## variable 9   0.044336971 -0.37380675 -0.0224249935
## variable 10  0.043215376 -0.37316095 -0.0162247651
## variable 11  0.043435543 -0.37309948 -0.0245954650
## variable 12  0.045416531 -0.37575287 -0.0200104086
## variable 13  0.045161134 -0.37621159 -0.0206043931
## variable 14  0.044602975 -0.37569136 -0.0144839228
## variable 15  0.002901088  0.02228831 -0.4056882953
## variable 16  0.003405222  0.02198276 -0.4068836318
```

```
## variable 17  0.003228802  0.02064986 -0.4107113728
## variable 18  0.002920826  0.02319232 -0.4056318015
## variable 19  0.002018519  0.02191736 -0.4087805942
## variable 20  0.001885975  0.02043594 -0.4078297610
```

```r
mod4$Tau
```

```
##             Component 1 Component 2 Component 3
## variable 1     4.349513    4.952224    4.961845
## variable 2     4.357094    4.951479    4.961185
## variable 3     4.348866    4.946652    4.955442
## variable 4     4.346970    4.942133    4.951918
## variable 5     4.346201    4.949399    4.957941
## variable 6     4.348419    4.945415    4.955085
## variable 7     4.353566    4.952076    4.961437
## variable 8     4.940075    4.341259    4.947637
## variable 9     4.943619    4.350731    4.950784
## variable 10    4.946272    4.354243    4.954138
## variable 11    4.941374    4.350682    4.947635
## variable 12    4.939956    4.342697    4.948068
## variable 13    4.934268    4.336865    4.942105
## variable 14    4.953804    4.353473    4.962576
## variable 15    4.963713    4.961313    4.266473
## variable 16    4.954953    4.952599    4.256391
## variable 17    4.956272    4.954215    4.246050
## variable 18    4.953989    4.951371    4.259347
## variable 19    4.962076    4.959717    4.256072
## variable 20    4.950147    4.948089    4.249971
```

`alphatau` and `betatau` can also be specified as D-dimensional array, in which case the Inverse Gamma will have component-specific hyperparameters:

$$\tau_{d,j} \sim IG(\alpha_d, \beta_d)$$

.

```r
# Set hyperparameter values
ctrl3 <- vbpca_control(center = FALSE, scalecorrection = -1,
                       plot.lowerbound = FALSE,
                       alphatau = c(.5, 50, 3), betatau = c(.5, .01, 10),
                       hypertype = 'component')


# Estimate the model
mod5 <- vbpca(X, D = 3, maxIter = 1e+03, priorvar = 'invgamma',
              control = ctrl3, verbose = FALSE )
```

```
## Warning: unscaled data - ELBO values might be positive.
```

```r
mod5$muW
```

```
##              Component 1  Component 2   Component 3
## variable 1  -0.378543478 -0.022403852  0.0025386295
## variable 2  -0.376019722 -0.021196983 -0.0088612115
## variable 3  -0.377066494 -0.022557126 -0.0057270031
## variable 4  -0.376774699 -0.022650241 -0.0008814042
## variable 5  -0.378720012 -0.022357681 -0.0078768015
## variable 6  -0.377102619 -0.019528818  0.0007032582
```

```
## variable 7  -0.376989874 -0.026248142  0.0014710887
## variable 8   0.021470521 -0.045359009 -0.0012343390
## variable 9   0.022018693 -0.049787441 -0.0042453561
## variable 10  0.020952041 -0.037610876  0.0018996414
## variable 11  0.021150966 -0.048838642 -0.0065159768
## variable 12  0.022990213 -0.048959831 -0.0017490218
## variable 13  0.022694280 -2.350959262 -0.0024775177
## variable 14  0.022197501 -0.048663411  0.0037657268
## variable 15  0.002988814  0.002048412 -0.4063509956
## variable 16  0.003472519 -0.001954198 -0.4074615231
## variable 17  0.003201167  0.008962963 -0.4112475488
## variable 18  0.003059452  0.006509927 -0.4063451138
## variable 19  0.002074293  0.003760222 -0.4093932228
## variable 20  0.001853458 -0.003935007 -0.4083201342
```

```r
mod5$Tau
```

```
##             Component 1 Component 2 Component 3
## variable 1     1.734909  4885.51101   0.3498302
## variable 2     1.737796  4897.91867   0.3498294
## variable 3     1.734283  4883.57903   0.3498025
## variable 4     1.733024  4882.65984   0.3497829
## variable 5     1.733356  4885.97064   0.3498127
## variable 6     1.733921  4914.16337   0.3497986
## variable 7     1.736550  4841.25146   0.3498290
## variable 8     1.974069  4535.17321   0.3497716
## variable 9     1.975981  4450.34973   0.3497902
## variable 10    1.977182  4671.07826   0.3498009
## variable 11    1.974679  4468.90214   0.3497765
## variable 12    1.974249  4466.51594   0.3497748
## variable 13    1.971273    18.20417   0.3497460
## variable 14    1.981100  4472.66011   0.3498376
## variable 15    1.982219  5007.34354   0.3469767
## variable 16    1.977733  5007.36672   0.3469196
## variable 17    1.978365  4988.53185   0.3468718
## variable 18    1.977266  4997.70264   0.3469310
## variable 19    1.981324  5004.97797   0.3469253
## variable 20    1.975283  5004.39805   0.3468847
```

Notice the different level of regularization obtained across the different components. In order to activate these 'component-specific' hyperpriors, `hypertype = 'component'` was specified.

Random `tau`, random `betatau`

It is also possible to specify a Gamma hyperprior on β (while α remains fixed):

$$\beta \sim Ga(\gamma, \delta).$$

This is achievable by setting `gammatau` (and `deltatau`) larger than 0 in the control parameters:

```r
# Specify component-specific Gamma(.01, 10) hyperpriors on betatau
ctrl4 <- vbpca_control(center = FALSE, scalecorrection = -1,
                       plot.lowerbound = FALSE,
                       alphatau = 1, betatau = 1,
                       gammatau = .01, deltatau = 10,
                       hypertype = 'component')
```

```r
# Estimate the model
mod6 <- vbpca(X, D = 3, maxIter = 1e+03, priorvar = 'invgamma',
              control = ctrl4, verbose = FALSE )
```

```
## Warning: unscaled data - ELBO values might be positive.
```

```r
mod6$muW
```

```
##              Component 1 Component 2  Component 3
## variable 1  -0.376611438 -0.04414830 -0.001252671
## variable 2  -0.373522305 -0.04527018 -0.009487805
## variable 3  -0.375303126 -0.04330237 -0.006687095
## variable 4  -0.374545174 -0.04476030 -0.003779739
## variable 5  -0.377135323 -0.04290004 -0.008530456
## variable 6  -0.375471472 -0.04449078 -0.002519439
## variable 7  -0.374841518 -0.04378470 -0.001756540
## variable 8   0.044091734 -0.37646021 -0.020075478
## variable 9   0.044279171 -0.37340282 -0.022181983
## variable 10  0.043230863 -0.37275698 -0.017353060
## variable 11  0.043633682 -0.37285414 -0.022657108
## variable 12  0.045225592 -0.37597481 -0.020153766
## variable 13  0.045217385 -0.37690958 -0.020321214
## variable 14  0.044311525 -0.37546975 -0.014960798
## variable 15  0.002829728  0.02204140 -0.405116611
## variable 16  0.003208895  0.02187057 -0.407344521
## variable 17  0.003155312  0.02095411 -0.410909109
## variable 18  0.002824576  0.02288662 -0.405229626
## variable 19  0.002136107  0.02174443 -0.408747297
## variable 20  0.002177548  0.02077097 -0.407952211
```

```r
mod6$Tau
```

```
##             Component 1 Component 2 Component 3
## variable 1     14.89847    49.95597    64.79114
## variable 2     15.06380    49.85157    64.48104
## variable 3     14.87935    49.45351    63.73992
## variable 4     14.89953    49.19672    63.74737
## variable 5     14.83146    49.84022    64.32954
## variable 6     14.87624    49.45017    63.91513
## variable 7     14.97986    49.94842    64.56200
## variable 8     48.95314    14.77489    62.76748
## variable 9     49.16960    14.97374    62.96668
## variable 10    49.41724    15.02556    63.45826
## variable 11    49.04970    14.98706    62.69018
## variable 12    48.92027    14.79264    62.77075
## variable 13    48.53974    14.68576    62.16584
## variable 14    50.07132    15.00763    64.71429
## variable 15    51.73889    51.41043    14.14011
## variable 16    51.12230    50.69433    13.94199
## variable 17    51.27494    50.94174    13.77126
## variable 18    51.01890    50.58728    14.03940
## variable 19    51.71281    51.34808    13.93690
## variable 20    50.73849    50.38297    13.85877
```

The posterior means of β can be accessed via

```
mod6$priorBeta
```

```
##            [,1]       [,2]       [,3]
## [1,] 0.02643281 0.02638618 0.02076472
## attr(,"names")
## [1] "beta 1" "beta 2" "beta 3"
```

`hypertype` specify the type of hyperprior for `beta`:

- `'common'` implies $\beta \sim Ga(\alpha, \beta)$;
- `'component'` implies $\beta_d \sim Ga(\alpha_d, \beta_d)$;
- `'local'` implies $\beta_{dj} \sim Ga(\alpha_{dj}, \beta_{dj})$.

Similar to `alphatau` and `betatau`, `gammatau` and `deltatau` can also be D-dimensional arrays for component-specific hyperpriors on β.

Global prior variances

So far, the parameter `global.var` has always ben set to `FALSE`, implying

$$w_{j,d} \sim N(0, \tau_{j,d}).$$

Setting `global.var = TRUE` will modify this formulation, which will switch to

$$w_{j,d} \sim N(0, \tau_d)$$

that is, component-specific variances (called 'global variances' in `vbpca`) will be estimated instead:

```
# Fixed prior global variances, updated via Type-II maximum likelihood:
mod7 <- vbpca(X, D = 3, maxIter = 1e+03, priorvar = 'fixed',
              updatetau = TRUE, control = ctrl, verbose = FALSE,
              global.var = TRUE)
```

```
## Warning: unscaled data - ELBO values might be positive.
```

```
mod7$muW
```

```
##              Component 1 Component 2   Component 3
## variable 1  -0.376586344 -0.04416414  0.0003398280
## variable 2  -0.373936445 -0.04582246 -0.0111461792
## variable 3  -0.375145315 -0.04305763 -0.0078812187
## variable 4  -0.374766738 -0.04473002 -0.0031117179
## variable 5  -0.376804669 -0.04285697 -0.0100225682
## variable 6  -0.375110723 -0.04446231 -0.0015008380
## variable 7  -0.375066004 -0.04363985 -0.0007179347
## variable 8   0.043915682 -0.37609858 -0.0194939220
## variable 9   0.044338601 -0.37381868 -0.0224109636
## variable 10  0.043215853 -0.37318635 -0.0161925187
## variable 11  0.043432402 -0.37310269 -0.0246469040
## variable 12  0.045419753 -0.37573441 -0.0200022167
## variable 13  0.045157688 -0.37615568 -0.0206098160
## variable 14  0.044605253 -0.37570521 -0.0144801414
## variable 15  0.002905193  0.02229189 -0.4056448021
## variable 16  0.003409730  0.02199104 -0.4067867230
## variable 17  0.003232816  0.02063848 -0.4105969740
## variable 18  0.002919683  0.02319284 -0.4055773542
## variable 19  0.002019241  0.02192068 -0.4087004821
## variable 20  0.001874190  0.02043083 -0.4077291009
```

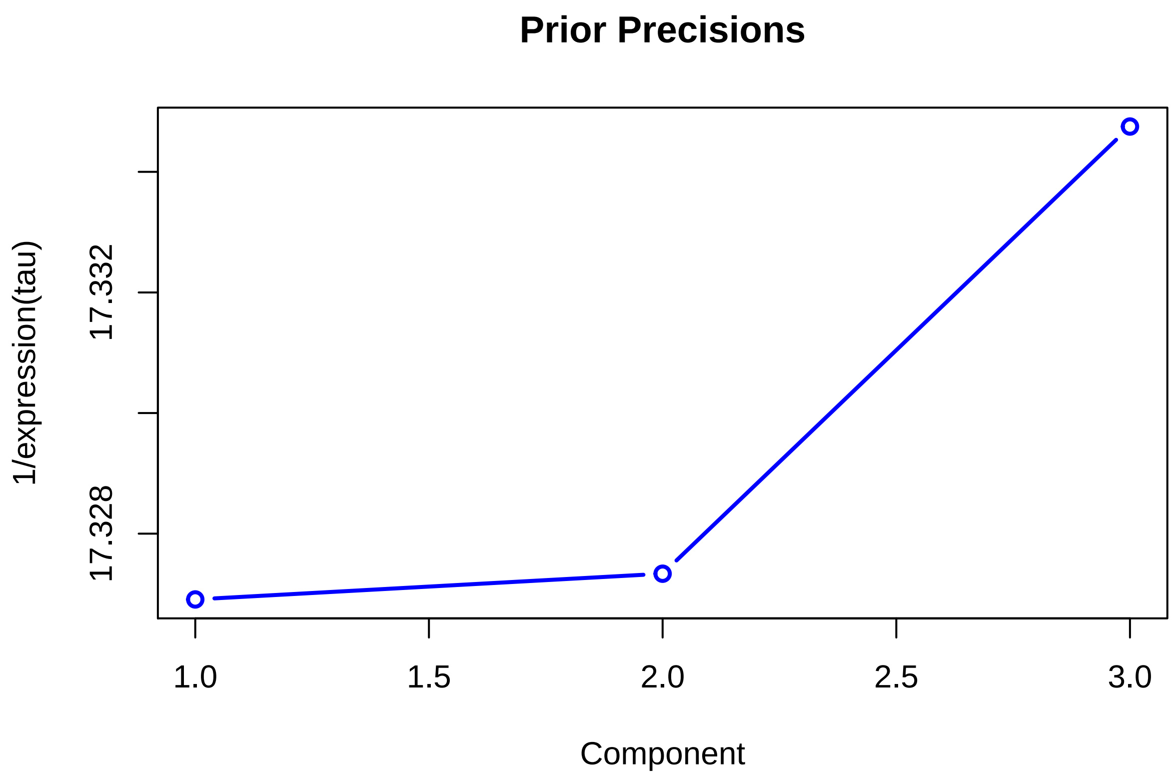

Figure 1: Prior variances for the first 3 components.

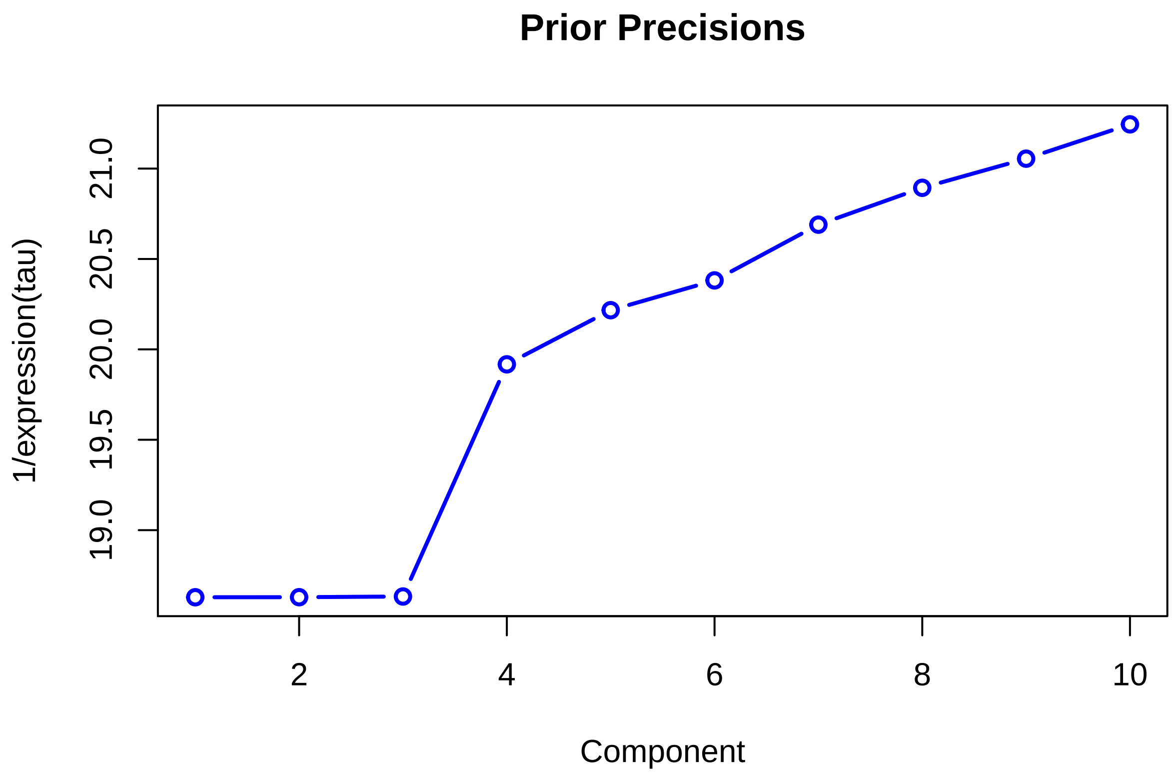

Figure 2: Scree-plot for 10 components.

```
mod7$Tau
```

```
## [1] 17.32691 17.32734 17.33475
```

Notice the plot of the precisions that appears in this case. This is useful when the number of components supported by the data is uncertain (scree-plot - see Figure 2):

```
mod8 <- vbpca(X, D = 10, maxIter = 1e+03, priorvar = 'fixed',
              updatetau = TRUE, global.var = TRUE,
              control = ctrl, verbose = FALSE )
```

```
## Warning: unscaled data - ELBO values might be positive.
```

Stochastic Search Variable Selection

By requiring `SVS = TRUE`, the model activates stochastic-search-variable-selection, a method described by George ad McCulloch (1993) for the Gibbs Sampler. The method has been adapted in *bayespca* for the Variational Bayes algorithm. The assumed 'spike-and-slab' prior for the (j, d)-th element of W becomes:

$$w_{j,d} \sim N(0, \pi\tau + (1 - \pi)\tau v_0)$$

where v_0 is a scalar which rescales the spike variance to a value close to 0. For this reason, v_0 should be a number included in $(0, 1)$, as close as possible to 0. π represents the prior probability of inclusion of the

j-th variable in the d-th component of the model. `vbpca` estimates the posterior probabilities of inclusion, conditional on X and the values in W.

While v_0 should be a small value close to 0, too small values of such parameter will shrink the variances τ too much, and no variable will eventually be included in the model. On the other hand, using a too large value for v_0 will not shrink the variances enough, and all posterior inclusion probabilities will be close to 1. v_0 should then be set with a grain of salt. Preliminary results from partial simulation studies have shown that values between 0.0001 and 0.005 lead to acceptable results, but adequate values of v_0 can be dataset-specific. Simulation studies have also shown that the method works better when Gamma priors are specified for τ.

In `vbpca`, the parameter v_0 is called `v0` in the control parameters of `vbpca_control`, while the prior inclusion probability is called `priorInclusion`. `priorInclusion` can be fixed, or assigned to a Beta hyperprior:

- among the control parameters of `vbpca_control`, set `beta1pi` smaller than or equal to 0 for fixed π;
- last, set `beta1pi` larger than 0 for Beta specifications.

When `beta1pi` is larger than 0, a Beta prior is assumed for π:

$$\pi \sim Beta(\beta_1, \beta_2).$$

In `vbpca`, $\beta 1$ can be controlled with the `beta1pi` argument and $\beta 2$ with the `beta2pi` argument in `vbpca_control`.

```
# SVS, fixed priorInclusion and InverseGamma(5, 1) for tau, v0 = .005
ctrl5 <- vbpca_control(center = FALSE, scalecorrection = -1,
                       plot.lowerbound = FALSE,
                       alphatau = 5, betatau = 1,
                       beta1pi = -1, v0 = 5e-03)
# Estimate the model with priorInclusion = 0.5
mod9 <- vbpca(X, D = 3, maxIter = 1e+03, priorvar = 'invgamma',
              SVS = TRUE, priorInclusion = 0.5, control = ctrl5,
              verbose = FALSE )
```

```
## Warning: unscaled data - ELBO values might be positive.
```

```
mod9$muW
```

```
##              Component 1 Component 2  Component 3
## variable 1  -0.376750692 -0.04448495 -0.004123683
## variable 2  -0.372459887 -0.04361185 -0.005752993
## variable 3  -0.375671975 -0.04356723 -0.004989270
## variable 4  -0.373836916 -0.04448097 -0.004703139
## variable 5  -0.377543656 -0.04319604 -0.005454948
## variable 6  -0.375846023 -0.04472475 -0.004528453
## variable 7  -0.375420482 -0.04399690 -0.004254785
## variable 8   0.044408056 -0.37704714 -0.019871878
## variable 9   0.043996739 -0.37297274 -0.020136604
## variable 10  0.043266209 -0.37568704 -0.019144528
## variable 11  0.043855992 -0.37127741 -0.019779142
## variable 12  0.044744851 -0.37518584 -0.019753061
## variable 13  0.045169754 -0.37455442 -0.019722877
## variable 14  0.043743494 -0.37734332 -0.018263649
## variable 15  0.002622365  0.02125415 -0.405655204
## variable 16  0.002687542  0.02139614 -0.407985339
## variable 17  0.002847711  0.02135195 -0.409082736
## variable 18  0.002536920  0.02132424 -0.406350545
## variable 19  0.002543555  0.02126640 -0.408688107
```

```
## variable 20  0.002678762  0.02114151 -0.408155324
```

```r
# SVS, priorInclusion with Beta(1,1) priors and InverseGamma(5, 1) for tau, v0 = .005
ctrl6 <- vbpca_control(center = FALSE, scalecorrection = -1,
                       plot.lowerbound = FALSE, alphatau = 5,
                       betatau = 1, beta1pi = 1, beta2pi = 1,
                       v0 = 5e-03)

# Estimate the model
mod10 <- vbpca(X, D = 3, maxIter = 1e+03, priorvar = 'invgamma',
               SVS = TRUE, priorInclusion = 0.5, control = ctrl6,
               verbose = FALSE )
```

```
## Warning: unscaled data - ELBO values might be positive.
```

```r
mod10$muW
```

```
##              Component 1 Component 2  Component 3
## variable 1  -0.376821881 -0.04443487 -0.004109436
## variable 2  -0.372330173 -0.04358419 -0.005697493
## variable 3  -0.375718284 -0.04343594 -0.004948158
## variable 4  -0.373892738 -0.04418224 -0.004672895
## variable 5  -0.377498804 -0.04329720 -0.005404988
## variable 6  -0.375807899 -0.04448271 -0.004504242
## variable 7  -0.375543840 -0.04396962 -0.004236982
## variable 8   0.044271834 -0.37708761 -0.019690504
## variable 9   0.043950285 -0.37295109 -0.019942379
## variable 10  0.043385056 -0.37575936 -0.018993345
## variable 11  0.043789968 -0.37125713 -0.019594610
## variable 12  0.044486008 -0.37516550 -0.019574991
## variable 13  0.044629057 -0.37452510 -0.019541734
## variable 14  0.043968445 -0.37746585 -0.018157271
## variable 15  0.002608161  0.02107559 -0.405655041
## variable 16  0.002664430  0.02120244 -0.407997729
## variable 17  0.002842477  0.02116040 -0.409084353
## variable 18  0.002533751  0.02111718 -0.406361686
## variable 19  0.002532574  0.02108326 -0.408694905
## variable 20  0.002668043  0.02094085 -0.408182161
```

The estimated posterior inclusion probabilities for the two models:

```r
mod9$inclusionProbabilities
```

```
##             Component 1 Component 2 Component 3
## variable 1   1.00000000   0.2120296  0.09726907
## variable 2   1.00000000   0.2068326  0.09851144
## variable 3   1.00000000   0.2084065  0.09897047
## variable 4   1.00000000   0.2139049  0.09843787
## variable 5   1.00000000   0.2044355  0.09823061
## variable 6   1.00000000   0.2146309  0.09848977
## variable 7   1.00000000   0.2091004  0.09793821
## variable 8   0.21428441   1.0000000  0.11723947
## variable 9   0.21111001   1.0000000  0.11752795
## variable 10  0.20612101   1.0000000  0.11542999
## variable 11  0.21070545   1.0000000  0.11707616
## variable 12  0.21594418   1.0000000  0.11702863
## variable 13  0.21982827   1.0000000  0.11758358
```

```
## variable 14  0.20661409   1.0000000  0.11227938
## variable 15  0.09660274   0.1179197  1.00000000
## variable 16  0.09702372   0.1193704  1.00000000
## variable 17  0.09668529   0.1185038  1.00000000
## variable 18  0.09705420   0.1192275  1.00000000
## variable 19  0.09635180   0.1178253  1.00000000
## variable 20  0.09741905   0.1191173  1.00000000
```

```r
mod10$inclusionProbabilities
```

```
##             Component 1 Component 2 Component 3
## variable 1   1.00000000  0.14850584  0.06710159
## variable 2   1.00000000  0.14481111  0.06794426
## variable 3   1.00000000  0.14528607  0.06824230
## variable 4   1.00000000  0.14842819  0.06788528
## variable 5   1.00000000  0.14366894  0.06775190
## variable 6   1.00000000  0.14930551  0.06791993
## variable 7   1.00000000  0.14647800  0.06755126
## variable 8   0.14943035  1.00000000  0.08087388
## variable 9   0.14764427  1.00000000  0.08106952
## variable 10  0.14487228  1.00000000  0.07966569
## variable 11  0.14722900  1.00000000  0.08075690
## variable 12  0.15003021  1.00000000  0.08073007
## variable 13  0.15126284  1.00000000  0.08108863
## variable 14  0.14596444  1.00000000  0.07756885
## variable 15  0.06663148  0.08142293  1.00000000
## variable 16  0.06690799  0.08237169  1.00000000
## variable 17  0.06668375  0.08179149  1.00000000
## variable 18  0.06693031  0.08225103  1.00000000
## variable 19  0.06646156  0.08135076  1.00000000
## variable 20  0.06717262  0.08216783  1.00000000
```

It is also possible to compare the (known) variable inclusion matrix vs. the estimated ones graphically. Let's plot a heatmap of such probabilities for model mod9:

```r
trueInclusions <- matrix(0, J, 3)
trueInclusions[1:7, 1] <- 1
trueInclusions[8:14, 2] <- 1
trueInclusions[15:20, 3] <- 1

par(mfrow=c(1,2))
image(1:ncol(trueInclusions), 1:nrow(trueInclusions),
        t(trueInclusions[J:1, ]), ylab = "", axes = FALSE,
        main = "True Inclusions", xlab = "",
        col = RColorBrewer::brewer.pal(9, "Blues"))
axis(side = 1, at = 1:3, labels = paste("Component ", 1:3 ))
axis(side = 2, at = 1:20, labels = paste("Var ", J:1 ))

fields::image.plot(1:ncol(trueInclusions), 1:nrow(trueInclusions),
                   t(mod9$inclusionProbabilities[J:1, ]), ylab = "", axes = FALSE,
                   main = "Estimated Inclusions", xlab = "",
                   col = RColorBrewer::brewer.pal(9, "Blues"))
axis(side = 1, at = 1:3, labels = paste("Component ", 1:3 ))
```

We can observe the estimated prior inclusion probabilities for mod10:

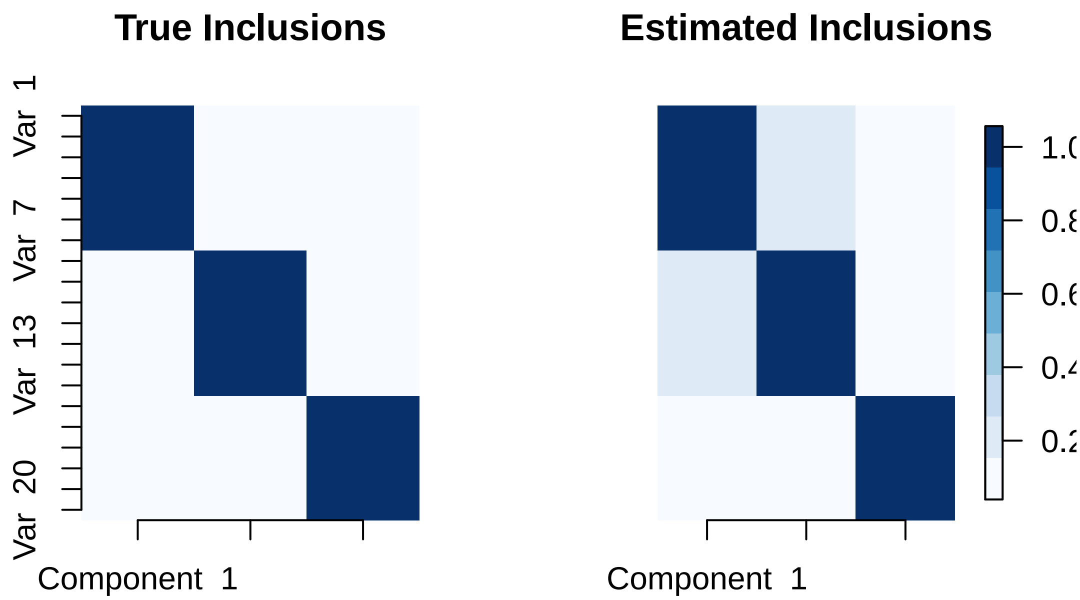

Figure 3: True and Estimated inclusion probabilities.

```
mod10$priorInclusion
```

```
##          [,1]
## [1,] 0.403084
## [2,] 0.403084
## [3,] 0.403084
```

Similar to the hyperparameters of the Inverse Gamma priors on τ, priorInclusion, beta1pi and beta2pi can also be specified as D-dimensional arrays. This will allow estimating the inclusion probabilities with different degrees of 'sparsity' for each component. For Beta priors, all elements of beta1pi must be larger than 0. Let us look at one example:

```
# Beta priors with different degrees of sparsity for each component
ctrl7 <- vbpca_control(center = FALSE, scalecorrection = -1,
                       plot.lowerbound = FALSE,
                       alphatau = 5, betatau = 1,
                       beta1pi = c(0.01, 1, 10), beta2pi = 1,
                       v0 = 5e-03)
# Estimate the model
mod11 <- vbpca(X, D = 3, maxIter = 1e+03, priorvar = 'invgamma', SVS = TRUE,
               priorInclusion = rep(0.5, 3), control = ctrl7, verbose = FALSE )
```

```
## Warning: unscaled data - ELBO values might be positive.
```

```
mod11$muW
```

```
##              Component 1 Component 2  Component 3
## variable 1  -0.376821889 -0.04445558 -0.004069637
```

```
## variable 2  -0.372328317 -0.04356824 -0.005747755
## variable 3  -0.375718432 -0.04342297 -0.004969024
## variable 4  -0.373892106 -0.04421774 -0.004668224
## variable 5  -0.377498784 -0.04325308 -0.005445156
## variable 6  -0.375807735 -0.04452319 -0.004486450
## variable 7  -0.375544189 -0.04396607 -0.004205692
## variable 8   0.044274077 -0.37707952 -0.019733044
## variable 9   0.043952934 -0.37296945 -0.020014261
## variable 10  0.043388589 -0.37572640 -0.018975238
## variable 11  0.043792754 -0.37126675 -0.019647359
## variable 12  0.044488033 -0.37518311 -0.019611465
## variable 13  0.044630711 -0.37454998 -0.019583937
## variable 14  0.043971729 -0.37742821 -0.018041956
## variable 15  0.002608500  0.02109164 -0.405680986
## variable 16  0.002664754  0.02122263 -0.407992055
## variable 17  0.002842888  0.02117744 -0.409091854
## variable 18  0.002534089  0.02114126 -0.406373044
## variable 19  0.002532960  0.02109991 -0.408701060
## variable 20  0.002668495  0.02096197 -0.408133648
```

```r
mod11$priorInclusion
```

```
##           [,1]
## [1,] 0.4016964
## [2,] 0.4443481
## [3,] 0.5816448
```

```r
mod11$inclusionProbabilities
```

```
##             Component 1 Component 2 Component 3
## variable 1   1.00000000  0.17351540   0.1319171
## variable 2   1.00000000  0.16904717   0.1336219
## variable 3   1.00000000  0.16967154   0.1342983
## variable 4   1.00000000  0.17356094   0.1335524
## variable 5   1.00000000  0.16759290   0.1332418
## variable 6   1.00000000  0.17457980   0.1336274
## variable 7   1.00000000  0.17104055   0.1328529
## variable 8   0.14865160  1.00000000   0.1581165
## variable 9   0.14687665  1.00000000   0.1585037
## variable 10  0.14412244  1.00000000   0.1556274
## variable 11  0.14646367  1.00000000   0.1579171
## variable 12  0.14924803  1.00000000   0.1578362
## variable 13  0.15047223  1.00000000   0.1586403
## variable 14  0.14521005  1.00000000   0.1512370
## variable 15  0.06626748  0.09554240   1.0000000
## variable 16  0.06654233  0.09667582   1.0000000
## variable 17  0.06631940  0.09598264   1.0000000
## variable 18  0.06656449  0.09654206   1.0000000
## variable 19  0.06609852  0.09545868   1.0000000
## variable 20  0.06680541  0.09644723   1.0000000
```

High posterior density intervals

It is also possible to require the computation of high probability density intervals for the elements of W, which can then be plotted with the `plothpdi` function, which internally calls `ggplot2` functionalities. *Note*: when normalised weights are require from the corresponding `vbpca_control` argument, the posterior density

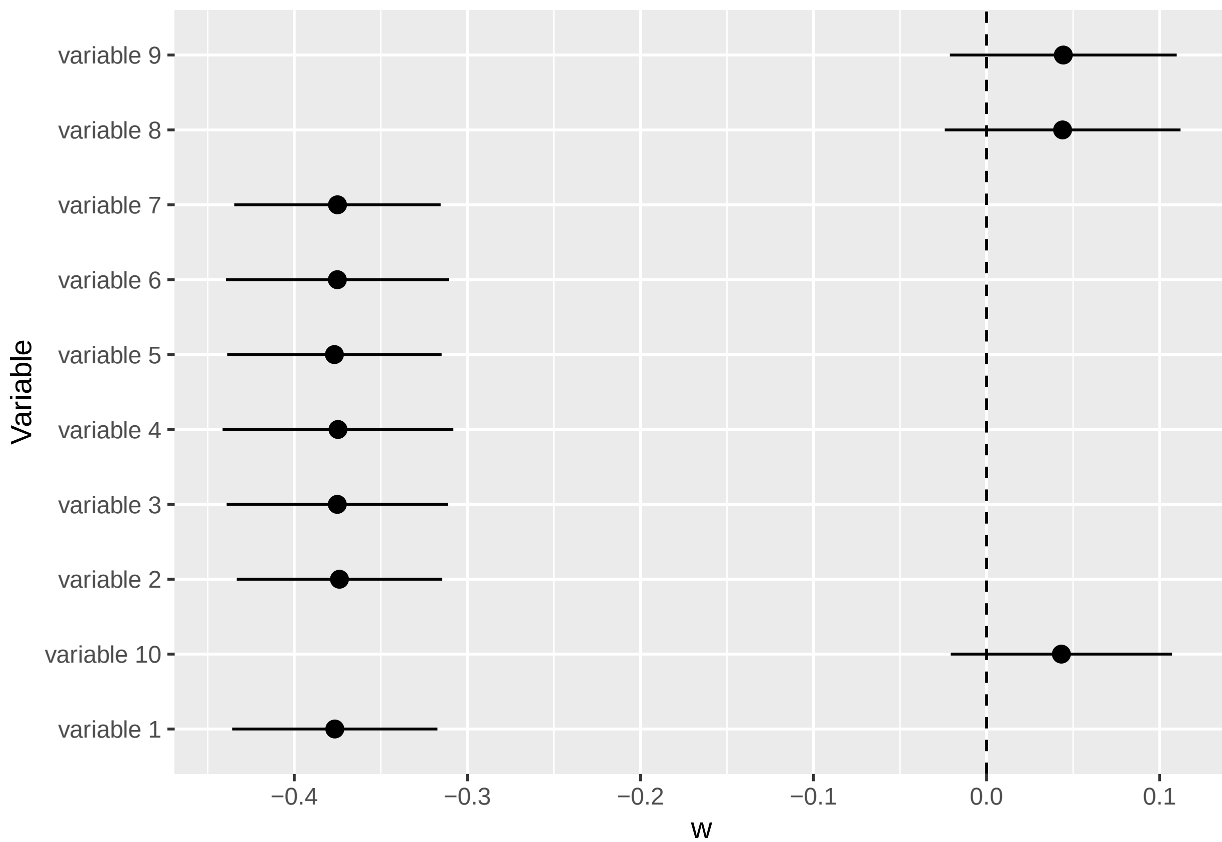

Figure 4: High posterior density intervals.

interval will still be returned in the original weights scale (thus, no normalisation is performed on the HPDIs).

```
# Set hyperparameter values and require 50% probability density intervals
ctrl8 <- vbpca_control(center = FALSE, scalecorrection = -1,
                       plot.lowerbound = FALSE,
                       alphatau = 2, betatau = .5,
                       hpdi = TRUE, probHPDI = 0.5)

# Estimate the model
mod12 <- vbpca(X, D = 3, maxIter = 1e+03, priorvar = 'invgamma',
              control = ctrl8, verbose = TRUE )
```

```
## Local prior variances : Inverse-Gamma, fixed hyperparameters.

## Warning: unscaled data - ELBO values might be positive.

## Iteration: 1 - ELBO: -2803.52
## Start # 1 has converged in 2 iterations; lower bound = -2802.98
```

```
# Plot HPD intervals for variables 1:10, component 1
plothpdi(mod12, d = 1, vars = 1:10)
```

Retrieve Principal Components

To compute the estimated components, simpy call:

```r
PCs <- X %*% mod1$muW
head(PCs, 15)
```

```
##       Component 1 Component 2 Component 3
##  [1,]   -59.19132  -78.592706  31.3401006
##  [2,]    28.97173 -118.789001 -29.0200757
##  [3,]   -11.00518   14.227039  -4.8429359
##  [4,]    92.16140  -33.606390 -28.1184464
##  [5,]   -41.61482 -212.440556  13.4800625
##  [6,]   113.51610  -20.107248   5.6778539
##  [7,]    98.45308  -73.892682  17.2711799
##  [8,]    42.05467 -142.922656 -68.0937444
##  [9,]   -57.38540  -66.586046  17.5396890
## [10,]    42.94090   51.286634  -0.2553017
## [11,]    36.39523  -11.871548  13.9383073
## [12,]   109.60474   -6.656482  25.3900540
## [13,]  -196.01791  110.020823  -9.5996904
## [14,]  -267.42318   71.336728  14.1676674
## [15,]    38.49334   22.034659 -32.6994037
```

References

1. C. M. Bishop. 'Variational PCA'. In Proc. Ninth Int. Conf. on Artificial Neural Networks. ICANN, 1999.
2. E. I. George, R. E. McCulloch (1993). 'Variable Selection via Gibbs Sampling'. Journal of the American Statistical Association (88), 881-889.