

The vabayelMix Package

February 16, 2008

Type Package

Title Variational Bayesian Mixture Modelling

Version 0.3

Date 2006-02-25

Author Andrew E Teschendorff

Maintainer <aet21@cam.ac.uk>

Depends mclust

Description Performs inference of a gaussian mixture model within a bayesian framework using an optimal separable approximation to the posterior density. The optimal posterior approximation is obtained using a variational approach.

License GPL version 2 or newer

R topics documented:

CostKL	2
MembProbFn2	3
UpdateCatw	4
UpdateMix	5
UseBasicPrior	6
pack	7
unbiasedKurt	8
vabayelMix	8

Index	11
--------------	-----------

 CostKL

Internal function for: Variational Bayesian Gaussian Mixture Model

Description

Computes the value of the cost function. Used for monitoring convergence.

Usage

```
CostKL(Ncat,data, m0, am0, aiv0, biv0, api0, m, am, aiv, biv, api, Catwm)
```

Arguments

`m0, am0, aiv0, biv0, api0` Prior hyperparameters, see `vabayelMix`

`m, am, aiv, biv` Posterior parameters.

`lambda` Categorical weight matrix, see References.

`s.lambda` Derived from `lambda`, see References.

Value

A list with the following components:

`mean` Means of gaussian posterior. Matrix of dimension `Ncat x Ndim`.

`ivarm` Inverse variances of gaussian posterior. Matrix of dimension `Ncat x Ndim`.

`ivara, ivarb` Parameters of gamma posterior. Matrices of dimension `Ncat x Ndim`.

`dapi` Parameters of dirichlet posterior giving weights of components.

Author(s)

Andrew Teschendorff<aet21@hutchison-mrc.cam.ac.uk>

References

- 1 D.J.MacKay: Developments in probabilistic modelling with neural networks-ensemble learning. In *Neural Networks: Artificial Intelligence and Industrial Applications*. Proceedings of the 3rd Annual Symposium on Neural Networks Nijmegen, Netherlands, Berlin Springer, 191-198 (1995).
- 2 J. W. Miskin : *Ensemble Learning for Independent Component Analysis*, PhD thesis University of Cambridge December 2000.
- 3 A. E. Teschendorff,...et al.: A variational bayesian mixture modelling framework for cluster analysis of gene expression data. Submitted to *Bioinformatics*.

`MembProbFn2`*Internal function for: Variational Bayesian Gaussian Mixture Model*

Description

Computes cluster membership probabilities of of samples.

Usage

```
MembProbFn2(data, NewVals, Nsamples)
```

Arguments

<code>data</code>	The data matrix
<code>NewVals</code>	Estimated parameter values
<code>Nsamples</code>	Number of samples

Value

A list with the following components:

<code>wc1</code>	Integer vector of length <code>Nsamples</code> specifying cluster membership of sample using maximum probability criterion
<code>probs</code>	Matrix of dimension <code>Ncat</code> x <code>Nsamples</code> giving cluster membership probabilities of samples.

Author(s)

Andrew Teschendorff<aet21@hutchison-mrc.cam.ac.uk>

References

- 1 D.J.MacKay: Developments in probabilistic modelling with neural networks-ensemble learning. In Neural Networks: Artificial Intelligence and Industrial Applications. Proceedings of the 3rd Annual Symposium on Neural Networks Nijmegen, Netherlands, Berlin Springer, 191-198 (1995).
- 2 J. W. Miskin : Ensemble Learning for Independent Component Analysis, PhD thesis University of Cambridge December 2000.
- 3 A. E. Teschendorff,...et al.: A variational bayesian mixture modelling framework for cluster analysis of gene expression data. Submitted to Bioinformatics.

UpdateCatw

Internal function for: Variational Bayesian Gaussian Mixture Model

Description

Updates categorical weights

Usage

```
UpdateCatw(Ncat, data, m, am, aiv, biv, api)
```

Arguments

`m, am, aiv, biv, api`
Posterior parameters.

Value

A list with the following components:

<code>cwm</code>	Categorical weight matrix. See References
<code>scw</code>	Derive from above. See References.

Author(s)

Andrew Teschendorff<act21@hutchison-mrc.cam.ac.uk>

References

- 1 D.J.MacKay: Developments in probabilistic modelling with neural networks-ensemble learning. In Neural Networks: Artificial Intelligence and Industrial Applications. Proceedings of the 3rd Annual Symposium on Neural Networks Nijmegen, Netherlands, Berlin Springer, 191-198 (1995).
- 2 J. W. Miskin : Ensemble Learning for Independent Component Analysis, PhD thesis University of Cambridge December 2000.
- 3 A. E. Teschendorff,...et al.: A variational bayesian mixture modelling framework for cluster analysis of gene expression data. Submitted to Bioinformatics.

UpdateMix

*Internal function for: Variational Bayesian Gaussian Mixture Model***Description**

Updates mean and variance parameters of mixture model.

Usage

```
UpdateMix(Ncat, data, m0, am0, aiv0, biv0, api0, m, am, aiv, biv, lambda, s.lambda)
```

Arguments

`m0, am0, aiv0, biv0, api0` Prior hyperparameters, see `vabayelMix`

`m, am, aiv, biv` Posterior parameters.

`lambda` Categorical weight matrix, see References.

`s.lambda` Derived from `lambda`, see References.

Value

A list with the following components:

`mean` Means of gaussian posterior. Matrix of dimension `Ncat x Ndim`.

`ivarm` Inverse variances of gaussian posterior. Matrix of dimension `Ncat x Ndim`.

`ivara, ivarb` Parameters of gamma posterior. Matrices of dimension `Ncat x Ndim`.

`dapi` Parameters of dirichlet posterior giving weights of components.

Author(s)

Andrew Teschendorff<aet21@hutchison-mrc.cam.ac.uk>

References

- 1 D.J.MacKay: Developments in probabilistic modelling with neural networks-ensemble learning. In Neural Networks: Artificial Intelligence and Industrial Applications. Proceedings of the 3rd Annual Symposium on Neural Networks Nijmegen, Netherlands, Berlin Springer, 191-198 (1995).
- 2 J. W. Miskin : Ensemble Learning for Independent Component Analysis, PhD thesis University of Cambridge December 2000.
- 3 A. E. Teschendorff,...et al.: A variational bayesian mixture modelling framework for cluster analysis of gene expression data. Submitted to Bioinformatics.

UseBasicPrior

Prior Function for Variational Gaussian Mixture Model

Description

This function implements an uninformative prior distribution for the cluster centers and variances, but allows the user to define prior weights for the clusters.

Usage

```
UseBasicPrior(data, weights.v)
```

Arguments

<code>data</code>	A matrix with columns representing variables and rows observations. Algorithm clusters observations.
<code>weights.v</code>	A vector of relative prior weights for the clusters.

Details

`weights.v` is a vector of length `Ncat`, the maximum number of clusters to look for.

Value

A list with following components. The first four are matrices of dimension `Ncat x Ndim`, `dapi` is a vector of length `Ncat`.

<code>mean</code>	the means of the cluster mean gaussian priors.
<code>varm</code>	the inverse variances for the cluster mean gaussian priors.
<code>ivara</code>	parameters for the gamma prior distribution of the inverse variances of the clusters. See references.
<code>ivarb</code>	parameters for the gamma prior distribution of the inverse variances of the clusters. See references.
<code>dapi</code>	weight vector specifying prior knowledge about the number of clusters.

Author(s)

Andrew Teschendorff<aet21@hutchison-mrc.cam.ac.uk>

References

- 1 D.J.MacKay: Developments in probabilistic modelling with neural networks-ensemble learning. In Neural Networks: Artificial Intelligence and Industrial Applications. Proceedings of the 3rd Annual Symposium on Neural Networks Nijmegen, Netherlands, Berlin Springer, 191-198 (1995).

- 2 J.W.Miskin : Ensemble Learning for Independent Component Analysis, PhD thesis University of Cambridge December 2000.
- 3 A. E. Teschendorff,...et al.: A variational bayesian mixture modelling framework for cluster analysis of gene expression data. Submitted to Bioinformatics.

pack

Profile analysis using Clustering and Kurtosis

Description

For a data matrix, selects features with kurtosis values in a specified range. Optionally, it further selects features based on whether their profiles are mixtures of gaussians or not.

Usage

```
pack(data.m, kurt.range=c(-2,0), cluster=T, method=c("bic", "vb"));
```

Arguments

<code>data.m</code>	Data matrix with features along the rows.
<code>kurt.range</code>	Desired range of kurtosis values.
<code>cluster</code>	Logical, to indicate whether additional cluster learning step is desired.
<code>method</code>	Character specifying model selection to be used (bic=EM-algorithm + BIC, vb=variational Bayesian + evidence bound).

Value

A list with the following components:

<code>out</code>	Matrix with rows labeling selected features and columns labeling kurtosis, cluster size and index position in data.m
<code>class</code>	A list with non-null elements giving the clustering classification of the selected features.

Author(s)

Andrew Teschendorff (aet21@hutchison-mrc.cam.ac.uk)

unbiasedKurt	<i>Internal function for: Variational Bayesian Gaussian Mixture Model</i>
--------------	---

Description

Gives unbiased kurtosis estimate for a sample vector.

Usage

```
unbiasedKurt(v)
```

Arguments

`v` A vector of values, missing entries are allowed

.

Value

A scalar giving the kurtosis.

Author(s)

Andrew Teschendorff (aet21@hutchison-mrc.cam.ac.uk)

vabaylMix	<i>Variational Bayesian Gaussian Mixture Model</i>
-----------	--

Description

Learns a gaussian mixture model from data using an optimal separable approximation to the posterior density. The optimisation uses a variational procedure and implements an iterative ensemble learning algorithm. The algorithm gives a framework in which to infer the number of clusters in the data set. Prior information may be incorporated through specification of hyperparameters in a prior distribution. Current version implements a gaussian mixture model where the covariances matrices are diagonal.

Arguments

`data` A matrix of dimension $N_s \times N_{dim}$ containing the data to be clustered. Algorithm clusters rows of matrix and treats columns as dimensions.

`prior` A list of various elements containing prior information as obtained for example by using `UseBasicPrior`. List elements are `prior$mean`, `prior$ivarm`, `prior$ivara`, `prior$ivarb` and `prior$dapi`. The first four are matrices of dimension $N_{cat} \times N_{dim}$, `prior$dapi` is a vector of length N_{cat} . `prior$mean` contains the means of the cluster mean gaussian priors. `prior$ivarm`

contains the inverse variances for the cluster mean gaussian priors. `prior$ivara` and `prior$ivarb` contain the parameters for the gamma prior distribution of the inverse variances of the clusters. `prior$dapi` is a weight vector specifying prior knowledge about the number of clusters. If `prior` is unspecified a complete uninformative prior is implemented that assumes rows to be mean normalised to zero.

<code>Ncat</code>	The maximum number of clusters or categories to look for in the data set. Algorithm switches off clusters it doesn't need. See References.
<code>nruns</code>	Number of ensemble learning optimisation runs to be performed. Each optimisation run uses a different (random) starting point.
<code>npick</code>	The <code>npick</code> runs (out of <code>nruns</code>) that best optimise the cost function. See References.
<code>MaxIt</code>	Maximum number of iterations to be performed for a single optimisation run.
<code>conv.tol</code>	Threshold tolerance level for establishing convergence of iterations.
<code>nCV</code>	Number of consecutive iterations to consider in establishing convergence of the run at level <code>conv.tol</code> .
<code>verbatim</code>	Logical. If true prints out estimates and cost function value per iteration.

Value

A list with the following components:

<code>estvals</code>	A list with components:
<code>mean</code>	Means of gaussian posterior. Matrix of dimension <code>Ncat x Ndim</code> . A row containing all zeros means that component is absent.
<code>ivarm</code>	Inverse variances of gaussian posterior. Matrix of dimension <code>Ncat x Ndim</code> .
<code>ivara,ivarb</code>	Parameters of gamma posterior. Matrices of dimension <code>Ncat x Ndim</code> .
<code>dapi</code>	Parameters of dirichlet posterior giving weights of components. A value of 1 means that component is absent.
<code>wcl</code>	A matrix of dimension <code>npick x Ns</code> . Each row gives cluster assignment of each row of data. Clusters are labeled by integers.
<code>probs</code>	A list of length <code>npick</code> , each list element is a matrix of dimension <code>Ns x Ncat</code> containing the probabilities of membership to clusters.
<code>costs</code>	A vector of length <code>nruns</code> specifying converged values of cost function.
<code>conv</code>	A binary vector of length <code>nruns</code> specifying if that run converged (0) or not (1).

Author(s)

Andrew Teschendorff<aet21@hutchison-mrc.cam.ac.uk>

References

- 1 D.J.MacKay: Developments in probabilistic modelling with neural networks-ensemble learning. In Neural Networks: Artificial Intelligence and Industrial Applications. Proceedings of the 3rd Annual Symposium on Neural Networks Nijmegen, Netherlands, Berlin Springer, 191-198 (1995).
- 2 J.W.Miskin : Ensemble Learning for Independent Component Analysis, PhD thesis University of Cambridge December 2000.
- 3 A. E. Teschendorff,...et al.: A variational bayesian mixture modelling framework for cluster analysis of gene expression data. Submitted to Bioinformatics.

Examples

```

NsTot <- 100;
Nspg <- 50;
Ng <- 2;
deg.idx <- 1 ;
data <- matrix( nrow=NsTot, ncol=Ng);
for( s in 1:Nspg ){
  data[s,] <- rnorm(Ng,0,0.25);
}
for( s in (Nspg+1):NsTot){
  data[s,] <- rnorm(Ng,0,0.25);
  data[s,deg.idx] <- rnorm(1,2,0.25);
}
types.idx <- c(rep(1,50),rep(2,50));
useprior.l <- UseBasicPrior(data,rep(1,4));
vbmix <- vabayelMix(data, prior=NA, Ncat=4, nruns=10, npick=2,MaxIt=500, conv.tol=0.001, nCV=
# or could use
# vbmix <- vabayelMix(data, prior=useprior.l, Ncat=4, nruns=10, npick=2,MaxIt=500, conv.tol=
plot(1:NsTot,vbmix$wcl[1,],type="h",col=types.idx);

```

Index

*Topic **cluster**

pack, [6](#)

UseBasicPrior, [5](#)

vabayelMix, [8](#)

*Topic **internal**

CostKL, [1](#)

MembProbFn2, [2](#)

unbiasedKurt, [7](#)

UpdateCatw, [3](#)

UpdateMix, [4](#)

CostKL, [1](#)

MembProbFn2, [2](#)

pack, [6](#)

unbiasedKurt, [7](#)

UpdateCatw, [3](#)

UpdateMix, [4](#)

UseBasicPrior, [5](#)

vabayelMix, [8](#)