Gérer des tableaux de données avec Matlab

D. Legland

18 mars 2011

Résumé

Quelques pistes de réflexion pour améliorer la gestion des tableaux de données numériques lorsque l'on travaille sous Matlab, ou un environnement équivalent.

Table des matières

1 Tableaux de données **2**

2 Gestion des données statistiques **2**
2.1 Niveaux de facteurs . . . 2
2.2 Stockage des facteurs . . . 2

3 Besoin fonctionnels **2**
3.1 Gestion de base . . . 2
Création . . . 2
Gestion des lignes . . . 3
Gestion des colonnes . . . 3
Gestion des données . . . 3
3.2 Entrées et sorties . . . 3
3.3 Gestion avancée . . . 3
3.4 Données . . . 3
Format de sauvegarde . . . 3
Divers . . . 4

4 Formats existants **4**
4.1 Format texte, CSV . . . 4
4.2 Format DIV . . . 4
4.3 Conventions de noms de fichiers . . . 4

1 Tableaux de données

Une grande majorité des résultats manipulés sont de nature numérique, et peuvent être organisés en tableaux de données.

Quelques outils d'importation et de traitement existent sous certains logiciels orientés statistiques (R, Excel), mais pour d'autres (tels que Matlab, Python) on est souvent limité. L'idée est de disposer d'une classe de haut niveau permettant d'effectuer rapidement des traitements complexes. Exemple :

```
tab.plot('var1', 'var2', 'r'); % affichage calibre
var1Mean = tab.mean('var1');    % moyenne d'une colonne
```

2 Gestion des données statistiques

En plus des données numériques, on manipule aussi des facteurs (variété, fruit, position, traitement...). Ces facteurs peuvent être représentés par des valeurs numériques entières, mais en pratique on a parfois du mal à se rappeler à quelle modalité appartient une valeur donnée. De plus, les légendes des graphiques ne sont pas très parlantes : variété 1 ou 2, position 4...

2.1 Niveaux de facteurs

Une possibilité de stockage est de garder une structure DIV, mais d'ajouter un champs 'l' (pour *level*), qui contient les niveaux de chaque facteur. Le champs l est un tableau de cellules avec autant d'éléments que de colonnes, et si une colonne i est considérée comme un facteur, la valeur de la cellule l{i} est un tableau de cellules contenant les noms des modalités associées à chaque niveau. On peut tester facilement si une colonne est un facteur en vérifiant si la cellule est vide ou non.

2.2 Stockage des facteurs

Les données sont représentées par un tableau de cellules, chaque cellule contenant une colonne, et chaque colonne étant stockée comme un tableau de valeurs numériques ou un tableau de cellules contenant des chaînes de caractères.

Solution plus lourde ?

3 Besoin fonctionnels

3.1 Gestion de base

Création

Table() crée une table à partir des données numériques fournies, et des éventuelles informations sur les lignes et les colonnes

Table.read() charge un fichier dans un objet tableau de données

Gestion des lignes

Gestions des lignes, et de leurs noms.

addRow
getRow
deleteRow
setRow
setRowName(i, str)
getRowName(i)
setRowNames(str{})
getRowNames

Gestion des colonnes

Gestions des colonnes, et de leurs noms.

addColumn
getColumn
deleteColumn
setColumn
setColumnName(i)
getColumnName(i)
setColumnNames
getColumnNames

Gestion des données

Une série de fonctions pour lire et modifier les données internes.

getValue(r, c)
setValue(r, c, v)

3.2 Entrées et sorties

Lire et écrire une table

read méthode statique qui retourne un tableau de données
write

3.3 Gestion avancée

3.4 Données

Format de sauvegarde

Chaque colonne est associée à une chaîne de caractères qui permet de contrôler la manière dont les nombre sont sauvés, afin de rendre les fichiers plus lisibles.

getFormat
setFormat
getFormats
setFormats

Divers

Pour changer le nom de la table, le fichier de sauvegarde, les commentaires (?)...

setName

getName

getFile

4 Formats existants

4.1 Format texte, CSV

Le format CSV est l'abbréviaton de comma separated values. Il contient donc des valeurs séparées par des virgules. On peut en général utiliser un autre séparateur que la virgule.

4.2 Format DIV

Le format DIV est un format CSV un peu normalisé, en particulier :

- la première ligne contient les noms des colonnes
- pour chaque ligne, le premier élément est le nom de l'individu, suivit de la liste des valeurs associées à chaque colonne.

Manques potentiels :

- le format de stockage des donnes (une chaîne %d ou %f) ?
- le nom de la table ?
- des commentaires ?

4.3 Conventions de noms de fichiers

Idée générale : nom composé de plusieurs morceaux, qui vont du général au particulier. Dans l'ordre :

- éventuellement : le type de mesure. Ex. : 'est' pour une estimation, 'meas' pour une mesure, 'map' pour une cartographie...
- les paramètres d'intérêt, qui caractérisent les colonnes. Ex. : 'Morpho', 'VSSv'...
- éventuellement, un caractère de soulignement suivi de l'unité choisie pour les individus. Ex. : '_slice', '_slab', '_fruit'...
- éventuellement, un point puis un mot-clé de concaténation de données. Ex. : '.fact' quand on rajoute des facteurs, '.co', '.acp', '.vp' quand on calcule une acp...