Instructions pour l'utilisation de PMANOVA 21 octobre 2003

Bill Shipley, Département de biologie, Université de Sherbrooke, Sherbrooke (Québec) J1K 2R1 CANADA.

Le logiciel PMANOVA effectue une analyse de la variance multivariée par permutations à un critère de classification avec contrôle sur différences entre blocs. La référence pour ce teste est (*1*) Il y a deux avantages à ce test par permutations : (1) nous ne sommes pas obligés de supposer une distribution normale multivariée avec une variance résiduelle constante entre les groupes, et (2) nous pouvons avoir plus de variables indépendantes que d'unités. Par contre, il existe toujours d'autres suppositions, comme pour un MANOVA typique. Ces suppositions sont :

1. Indépendance des unités expérimentales. Cette supposition sera atteinte dans vos expériences.
2. Homogénéité des variances. On suppose que la variance résiduelle (intra-groupe) de chacune de vos variables à l'intérieur des groupes (traitements) est égale. On peut accepter un certaine degré de déviation de cette supposition; la règle générale est que la ratio de variances intra-groupe maximale : minimale est moins que 10. Par contre, la violation de cette supposition pour les comparaisons de deux groupes à la fois est plus importante à cause d'une plus petite effectif.

La règle générale : si le teste PMANOVA pour tous les groupes ensemble n'est pas significatif, alors ne faites pas les testes 2 à 2.

Dispositif expérimental au laboratoire de Carole Beaulieu : Typiquement, vous divisez un champs en plusieurs « blocs ». Dans chaque bloc vous faites un nombre p de « parcelles ». À chaque parcelle vous attribuez un traitement différent (donc vous avez p traitements). Ce dispositif est un placement aléatoire avec blocs, sans réplication à l'intérieur des blocs. Vous prenez plusieurs échantillons de sol « carottes » par parcelle et 3 réplicats par carotte. Puisque le traitement est appliqué au niveau des parcelles et non à chaque carotte indépendamment, la parcelle définit l'unité d'observation (ou expérimentale) et non les carottes.

Définitions :

Unité d'observation ou unité expérimentale : Ceci est l'unité (la « chose ») sur laquelle vous avez appliqué les traitements expérimentaux. Par exemple, si vous avez appliqué un traitement à une parcelle de terrain, et ensuite pris plusieurs échantillons de cette parcelle, alors l'unité d'observation et la parcelle et non chaque échantillon.

Modèle :

Y = Bloc + Traitement + Erreur(parcelle).

Y = $Y_{1jk...}$, $Y_{2jk...}$, … $Y_{njk...}$ c'est à dire, un vecteur de plusieurs variables indépendantes.

Hypothèse nulle : $\overline{\boldsymbol{\mu}}_1 = \overline{\boldsymbol{\mu}}_2 = \ldots \overline{\boldsymbol{\mu}}_n$. Il n'y a pas de différences entre les moyennes des variables parmi les n groupes.

Procédure :

Une valeur de F de Fisher est calculée pour chaque variable indépendante, en corrigeant pour les différences entre blocs, et en tenant compte de la structure nichée des données (si nécessaire, par exemple si vous avez plusieurs carottes par parcelle). Ensuite, les parcelles sont allouées de façon aléatoire aux traitements à l'intérieur de chaque bloc par ordinateur pour créer une base de données où il n'y a aucune différence entre traitements, et les valeurs de F (appellons-le un « F* » sont encore calculées pour chaque variable indépendante. Ceci est répété un grand nombre (N) de fois (au moins 1000, mais plus si vous voulez plus de précision) pour créer un grand nombre de bases de données avec la même structure que vos données et la même distribution de probabilité, mais dans lesquelles l'hypothèse nulle que les moyennes

des traitements dans chaque bloc sont égales est vraie. Ensuite, le nombre de fois que les F* calculés dans ces bases de données sont supérieures au F dans les vraies données est calculé. La proportion « p » des fois ou le F* >F parmi les 1000 bases de données est une estimation de la probabilité d'avoir observé les différences entre traitements dans vos données purement par chance. Un intervalle de confiance à 95 % de ces probabilités estimées est $p \pm \sqrt{(p(1-p)/N)}$.
Pour tester l'hypothèse nulle, on combine les probabilités individuelles pour chaque variable comme suit :

$C = -2\sum_{i=1}^{k} Ln(p_i)$. Cette statistique est distribuée selon une distribution χ^2 avec 2k degrés de liberté selon l'hypothèse nulle; k est le nombre de variables.

Ensuite, vous pouvez tester l'égalité des moyennes pour seulement 2 groupes (traitements) à la fois. Il faut modifier le seuil de signification à α(n si vous faites ceci pour n pairs de groupes. NOTEZ QUE VOUS FAITE CES TESTS DE COMPARAISON DE DEUX GROUPES À LA FOIS SEULEMENT SI

Structure de la base de données :

1. Entrer vos données en EXCEL. Chaque ligne représente une observation multivariée (exemple, des mesures pour chaque source de carbone pour un échantillon de sol. Vous pouvez inclure plusieurs carottes par parcelle, même plusieurs réplicats par carotte, mais il faut l'indiquer dans la base de donnes. Il faut une colonne pour indiquer le bloc (un numéro) dans laquelle l'observation appartient, une colonne pour indiquer le traitement (un numéro) dans laquelle l'observation appartient, une colonne pour indiquer l'unité d'observation (un numéro) – c'est à dire l'unité sur laquelle vous avez appliqué le traitement (exemple : parcelle) et aussi des colonnes pour indiquer les sous-unités d'observations si vous avez plusieurs observations (exemple, carottes) par unité d'observation. Ensuite, il y a une colonne pour chaque variable indépendante (exemple, absorbance pour une source de carbone).
ON NE PEUT PAS AVOIR DES DONNÉES MANQUANTES. SI VOUS EN AVEZ, IL FAUT ENLEVER LA LIGNE AU COMPLET.

2. Sauver le fichier en format EXCEL et ensuite le sauver en format texte (.TXT), qui est un format ASCII.

3. Enlever la ligne (si nécessaire) avec les noms des colonnes; il ne faut pas avoir du texte dans le fichier texte seulement des chiffres. Notez la colonne qui contient l'appartenance aux blocs, aux traitements, aux unités d'observations et les colonnes contenant les variables. Notez également le nombre total de colonnes dans le fichier. Le logiciel vous demandera ces informations.

4. Garder le logiciel PMANOVA.EXE dans un sous-répertoire. Garder le fichier DOS4GW.EXE dans le même sous-repertoire; sinon PMANOVA ne marchera pas.

4. Ouvrir une fenêtre MS-DOS ou double cliquer sur le fichier PMANOVA.EXE et suivre les instructions.

Exemple d'une base de données

Voici une partie de la première ligne d'un fichier qui s'appelle sem3.dat. Cette ligne contient les résultats d'une des trois réplicats d'une carotte de sol prise dans la première parcelle (qui a reçu le traitement 1) du bloc 1. Les colonnes sont:

1 bloc
2 traitement
3 unité d'observation (parcelle)
4 carotte
5 réplicat
6 valeur d'absorbance pour la première source de carbone

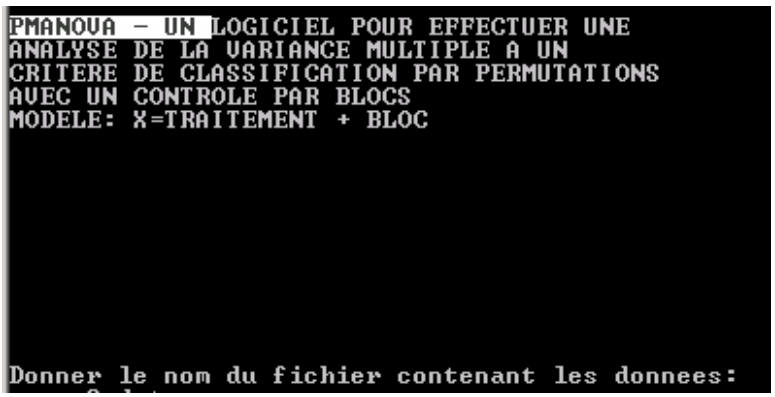
7 à 36 valeurs d'absorbance pour les autres sources de carbone.

1 1 1 1 1 1.38 0.845

Notez qu'il y a au moins une espace entre chaque chiffre et il y a aucun caractère autre que des chiffres.

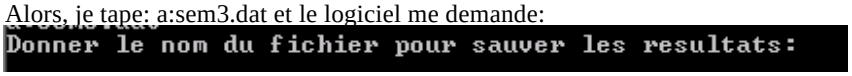
Exemple d'utilisation de PMANOVA

Après avoir double-cliqué sur PMANOVA.EXE, voici ce qui apparaît sur l'écran:

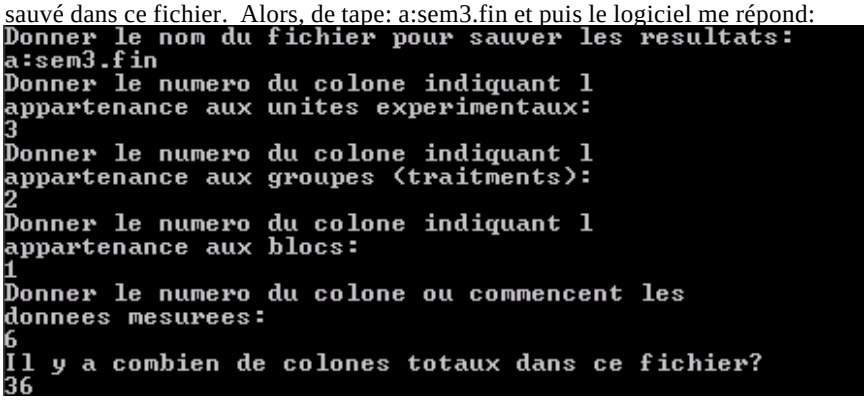


Il faut maintenant taper le nom compète du fichier texte contenant les données. Si ce fichier se trouve dans un sous-répertoire autre que le sous-répertoire dans lequel PMANOVA.EXE réside, il faut aussi donner la piste compète. Par exemple, si le fichier s'appelle "sem3.dat" et il se trouve sur une disquette dans "a", il faut taper: a:sem3.dat. Le règles pour les noms de fichiers sont les mêmes que pour DOS; pas plus que 8 caractères (n'incluant pas la piste), un point, et puis pas plus de 3 caractères.

Alors, je tape: a:sem3.dat et le logiciel me demande:



Ici, je donne n'importe quel nom de fichier, suivant les règles de DOS. Les résultats de l'analyse seront sauvé dans ce fichier. Alors, de tape: a:sem3.fin et puis le logiciel me répond:



Dans le fichier sem3.dat, la première colonne contient l'appartenance aux blocs, la deuxième colonne contient l'appartenance aux traitements et la troisième colonne contient l'appartenance aux parcelles (ici les unités expérimentales). Les colonnes 4 et 5 contiennent l'information sur les carottes de sols et les réplicats de chaque carotte, ce qui n'est pas nécessaire pour le logiciel. La première valeur d'absorbance (donc, la première variable indépendante) se trouve à la colonne 6 et les autres variables se trouvent aux colonnes 7,8, …36.

Le logiciel vous demande ensuite:

```
Voulez-vous voir les details sur les calculs
(Carrees moyennes, valeurs de F etc.) ou
simplement les resultats finaux?  Entrer le
numero 1 si vous voulez les details et entrer
le numero 0 si vous voulez simplement les
resultats finaux:
```

Si vous voulez voir la décomposition de la variance pour chaque variable et le F pour l'effet des traitements, alors taper 1, sinon 0. Puisqu'il y a 31 variables indépendantes dans ce fichier, nous allons choisir simplement les résultats finaux, alors je tape 0. Le logiciel répond:

```
  288 lignes lues du fichier
 Bloc   Nombre de lignes
   1        72
   2        72
   3        72
   4        72
 Groupe   Nombre de lignes
  1                    72
  2                    72
  3                    72
  4                    72
  Unite   Nombre de lignes
  1                    18
  2                    18
  3                    18
  4                    18



Combien de permutations desirez-vous?
Pour plus de precision, choisisez >5000
```

Vérifier si le nombre de lignes lues est correct; sinon il y a probablement des données manquantes dans votre base de données. Après avoir donné des informations sur le nombre de lignes comprises dans chaque bloc, groupe (traitement) et unité d'observation, le logiciel vous demande d'entrer le nombre de permutations indépendantes à utiliser pour calculer les probabilités. Normalement, 1000 est suffisante. Plus de permutations augmentent la précision des probabilités estimées mais augmente le temps pour faire l'analyse. Normalement 1000 permutations prendront environ 10 secondes et 5000 peut prendre jusqu'à 3 minutes. Nous allons choisir 1000 permutations.

```
    13            0.21800   0.02559
    14            0.67500   0.02903
    15            0.77100   0.02604
    16            0.36100   0.02977
    17            0.23000   0.02608
    18            0.64900   0.02958
    19            0.78100   0.02563
    20            0.96400   0.01155
    21            0.41800   0.03057
    22            0.50800   0.03099
    23            0.74200   0.02712
    24            0.85900   0.02157
    25            0.59200   0.03046
    26            0.73300   0.02742
    27            0.24800   0.02677
    28            0.79600   0.02498
    29            0.59700   0.03040
    30            0.97800   0.00909
    31            0.00700   0.00517
 Chi-carre est   45.43 avec  62 ddl
 Probabilite du chi-carre: 0.943504
Maintenant, voulez-vous comparer seulement
deux traitements (groupes) a la fois?
Repondre non pour quitter.
```

La probabilité (ligne 2) et ses intervalles de confiance à 95% (ligne 3) sont données pour chaque variable indépendante (ligne 1). Chaque probabilité se réfère à l'hypothèse nulle qu'il n'y aucune différence entres les traitements pour la variable indépendante en question. Ensuite la valeur du Chi carré est donnée avec ses degrés de liberté et ensuite la probabilité d'observer une telle valeur du Chi carré par chance avec vos données, en supposant qu'il n'y a aucune différence entre les traitements pour aucune variable indépendamment.
Maintenant, on peut terminer - en tapant non - ou continuer pour tester les différences entre des paires de groupes (traitements) - en tapant oui. Normalement, on choisira de quitter puisque nous savons déjà qu'il n'a pas de différences significatives entre les traitements, mais on va taper oui pour montrer comment comparer les traitements deux à la fois.

```
Donner le numero correspondant au premier
groupe (traitement)
1
Donner le numero correspondant au deuxieme
groupe (traitement)
2
```

Le logiciel vous demande de donner le numéro correspondant au premier groupe (par exemple, le groupe contrôle) et puis le numéro correspondant au deuxième groupe dans la comparaison. Notez que vous devez modifier votre seuil de signification (a) à a/n si vous faites n comparaisons.

Le logiciel vous pose les mêmes questions. Une fois que vous avez compété l'analysem quittez en tapant non à la question appropriée.

1. F. Pesarin, *Multivariate permutation tests with applications in biostatistics*. (Wiley, Chichester, 2001).