MATH 60602 *Analyse multidimensionnelle appliquée*

Examen de pratique

Questionnaire

Examinateur: Léo Belzile

Instructions: L'examen est d'une durée de 180 minutes. Aucune documentation écrite n'est permise. L'utilisation d'un ordinateur ou de tout autre matériel électronique est interdit. Une calculatrice non programmable est autorisée.

La répartition des 50 points de l'examen se trouve dans la marge de droite.

Les questions avec une étoile ($\star$) demandent une réponse plus élaborée. Prévoyez du temps en conséquence.

Vous devez rendre ce **questionnaire** et le feuillet avec les annexes à la fin de l'examen.

Aide-mémoire

- Propriétés de l'exponentielle: $\exp(a+b) = \exp(a)\exp(b)$, $\exp(0) = 1$
- Propriétés du logarithme népérien: $\ln(ab) = \ln(a) + \ln(b)$, $\ln(1) = 0$.
- sensibilité: $\Pr(\widehat{Y} = 1 \mid Y = 1)$, spécificité: $\Pr(\widehat{Y} = 0 \mid Y = 0)$, bonne classification: $\Pr(\widehat{Y} = Y)$.
- Critères d'information: $\mathsf{AIC} = -2\ell(\widehat{\boldsymbol{\theta}}) + 2p$ et $\mathsf{BIC} = -2\ell(\widehat{\boldsymbol{\theta}}) + \ln(n)p$ où $\widehat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell(\boldsymbol{\theta})$, $p = \dim(\boldsymbol{\theta})$.
- modèle de régression logistique: paramétrisation avec référence $Y = k$

$$\ln\left\{\frac{\Pr(Y=j \mid \mathbf{X})}{\Pr(Y=k \mid \mathbf{X})}\right\} = \eta_j = \beta_{0j} + \beta_{1j}\mathrm{X}_1 + \cdots + \beta_{jp}\mathrm{X}_p, \qquad j \neq k.$$

- modèle à risques proportionnels de Cox avec stratification pour X_1 catégorielle:

$$h(t;\mathbf{X}) = h_{\mathrm{X}_1}(t)\exp(\beta_2\mathrm{X}_2 + \cdots + \beta_p\mathrm{X}_p).$$

Section réservée au correcteur

Question:	1	2	3	4	5	Total
Points:	13	7	10	8	12	50
Score:						

Question 1. Durée des abonnements mensuels de transport en commun **13**

La Société des transports de Montréal (STM) a mandaté la firme Léger pour faire un sondage sur le comportement des usagers et usagères de transport en commun à Montréal. Les informations collectées lors de l'enquête sur $n = 1318$ personnes détentrices d'un abonnement annuel ou mensuel au cours des trois dernières années sont les suivantes:

- `clientele`: variable catégorielle, une parmi
  - jeune de 12 à 17 ou étudiant(e) de moins de 25 ans (`jeune`),
  - régulier (`regulier`) et
  - 65 ans et plus (`65+`).
- `activite`: variable binaire indiquant si la personne est employée ou aux études (`oui` ou `non`).
- `nmens`: nombre de passages mensuels moyens
- `nmoisabo`: nombre de mois avec un abonnement
- `abo_valide`: est-ce que l'abonnement était valide au moment de la fin de l'étude, 1 si oui, 0 sinon.

La STM s'intéresse à la durée des abonnements mensuels (`nmoisabo`) en fonction des informations sociodémographiques collectées. Certaines personnes sondées sont toujours titulaires d'un titre mensuel lors de l'enquête; il n'y a aucune perte de suivi.

1.1 Selon l'**Annexe 1**, combien de personnes parmi les 1318 détiennent toujours un abonnement à la fin de la période d'étude? [2]

> **Solution:** Pour déterminer le nombre d'observations censurées à droite, il faut regarder pour chaque groupe le nombre total d'événements. On obtient que $534 = 1318 - (78 + 184 + 522)$ personnes détiennent toujours un abonnement.

1.2 À l'aide des statistiques présentées dans l'**Annexe 1**, rapportez une estimation **fiable** de la moyenne et la médiane de la durée d'abonnement au sein de la population de 65 ans et plus détentrice d'abonnements mensuels. Si ce n'est pas possible, expliquez pourquoi. [2]

> **Solution:** Les statistiques descriptives usuelles ne prennent pas en compte la censure et donc sous-estiment le temps de détention d'un abonnement. Les statistiques descriptives sont 30.48 mois (moyenne) et 19 mois (médiane) sont donc obtenues à partir du Tableau A3. L'estimation de la moyenne est fiable puisque la plus grande valeur pour les 65 ans et plus n'est pas censurée (Figure A1).

1.3 Vous tracez les courbes de survie pour chaque clientèle; le résultat est présenté à la Figure A1 de l'**Annexe 1**. Identifiez la courbe correspondant à chaque clientèle et justifiez votre réponse. [2]

- `jeune`:
- `regulier`:
- `65+`:

Justifiez votre réponse.

> **Solution:** Le classemenent est `regulier` (A), `65+` (C) et `jeune` (B). Plusieurs justifications sont possibles: le Tableau A3 indique la survie maximale pour les jeunes (courbe B). On

peut également utiliser les valeurs maximales (censurées ou pas, Tableau A1) et les médianes des courbes de survie (Tableau A2).

1.4 À partir du Tableau A3 de l'**Annexe 1**, calculez le premier quartile du temps d'abonnement pour les **jeunes** (c'est-à-dire, le temps à partir duquel 25% des personnes sont désabonnés). [2]

Solution: 9 mois

Vos gestionnaires sont intéressés à savoir comment le profil de la clientèle affecte le temps de détention de laisser-passer mensuels. Vous ajustez un modèle de Cox avec `nmens` et `activite` en stratifiant selon la `clientele`. Les résultats sont fournis dans l'**Annexe 2**.

1.5 Interprétez le coefficient pour `nmens` (**en pourcentage d'augmentation ou de diminution** du risque). **Attention à votre interprétation** [2]

> **Solution:** *Ceteris paribus* (pour une même activité **et** une même clientèle), chaque augmentation du nombre de passage mensuels de 1 diminue le risque de désabonnement de 3.8% en moyenne.

1.6 (⋆) Quels sont les avantages et inconvénients de la stratification par `clientele`? Expliquez en quoi la sortie du modèle avec `clientele` comme variable explicative différerait de celle rapportée dans l'**Annexe 2**. [3]

> **Solution:** Si le postulat de risques proportionnels ne tient pas, les erreurs-types sont faussées.
>
> En stratifiant, on calcule le risque de base séparément pour chaque clientèle (`jeune`, `65+` et `regulier`) sans faire le postulat de risque proportionnels pour cette variable à priori.
>
> Le risque de base est estimé séparément, donc moins d'observations.
>
> En contrepartie, on n'a plus de coefficients pour la variable `clientele` et on ne peut tester l'effet de la variable explicative (on peut néanmoins comparer la survie).

Question 2. Analyse factorielle **7**

On vous demande de faire une analyse des résultats du sondage administré aux 1318 participant(e)s contacté(e)s par Léger. Toutes les questions sont mesurées sur des échelles de Likert de 1 à 10: où 1 indique très insatisfait et 10 très satisfait.

2.1 Quels sont les objectifs de l'analyse factorielle exploratoire? [2]

Solution:

- Modéliser la covariance/corrélation à l'aide d'un nombre réduit de paramètres
- Réduction de la dimension d'un ensemble de variables explicatives à l'aide de facteurs latents

2.2 L'**Annexe 3** contient les résultats d'un sondage de 13 questions et l'estimation du modèle d'analyse factorielle. Sur la base des sorties, choisissez un nombre adéquat de facteurs. Justifiez votre réponse. [3]

Solution: Selon les critères d'information, choisir quatre facteurs (BIC) ou cinq facteurs (AIC, test d'hypothèse comparant corrélation empirique et modèle ajusté) serait adéquat pour représenter la matrice de corrélation.

En revanche, la solution avec cinq regroupements ne donne aucune variable corrélée avec le cinquième facteur: tous les chargements sont inférieurs à 0.3. Donc quatre facteurs ici

2.3 À partir du modèle avec le nombre de facteurs choisi, fournissez une typologie des facteurs. [2]

Solution: Plusieurs descriptions possibles

- F1: offre en transport
- F2: service et installations
- F3: service et sécurité
- F4: offre des titres

Question 3. Données manquantes dans une analyse de regroupements **10**

On veut segmenter une base de données pour faire une analyse de regroupements, mais cette dernière contient des valeurs manquantes. Pour pallier à ce problème, on fait de l'imputation multiple et on crée 100 copies de la base de données avec les valeurs manquantes imputées. On procède ensuite à une analyse de regroupements avec la méthode des K-moyennes séparément pour chaque copie imputée de la base de données.

3.1 ($\star$) Quelles sont les principales étapes d'une analyse de regroupement? [3]

Solution:

- Choisir les variables pertinentes à l'analyse. Cette étape peut nécessiter de créer, transformer de nouvelles variables ou d'aggréger les données.
- Choisir les hyperparamètres de l'algorithme (nombre de regroupements, valeurs initiales) et la mesure de dissemblance.
- Valider la qualité de la segmentation (interprétabilité, taille des groupes, homogénéité des regroupements).
- Interpréter les regroupements obtenus à partir des prototypes (barycentres)

3.2 Suggérez un **critère** pour **choisir** le nombre de regroupements avec la méthode des K-moyennes. Expliquez ce dernier brièvement. [2]

Solution: Le critère du R^2 (critère du coude) ou le R^2 semi-partiel. Le $R^2 = 1 - \mathsf{SCD}_K/\mathsf{SCD}_1$ représente la somme du carré des distances intra-groupe vs somme du carré par rapport à la moyenne globale. On cherche le coude (un point d'inflexion où l'amélioration n'est plus notable).

3.3 Expliquez comment vous adapteriez ce dernier pour choisir un seul nombre de groupes k à partir des résultats pour les 100 imputations. [2]

Solution: Dans les deux cas, on sauvegarderait la valeur du critère pour $k = 1, \ldots, 10$, par exemple.

Ensuite, on peut calculer la moyenne (ou tracer une boîte à moustaches pour chaque valeur de k) et l'écart-type pour sélectionner le modèle adéquat à l'aide d'un seul graphique.

3.4 La Figure 1 montre le résultat de l'analyse de regroupements avec $k = 3$ groupes pour deux des 100 bases de données imputées. Elle illustre le problème de **permutation des étiquettes**. Suggérez une approche pour combiner les résultats des 100 analyses afin d'avoir une seule segmentation des sujets. *Indication: la sortie du modèle est une étiquette de groupe ($G_1, G_2, \ldots, G_k$) pour chaque copie de la base de données.* [3]

Solution: Toute réponse logique serait acceptée. Par exemple, on pourrait réassigner les numéros selonle barycentre pour s'assurer qu'ils coïncident (en ordre selon les coordonnées initiales ou les composantes principales).

Par la suite, il suffit d'assigner à la classe majoritaire.

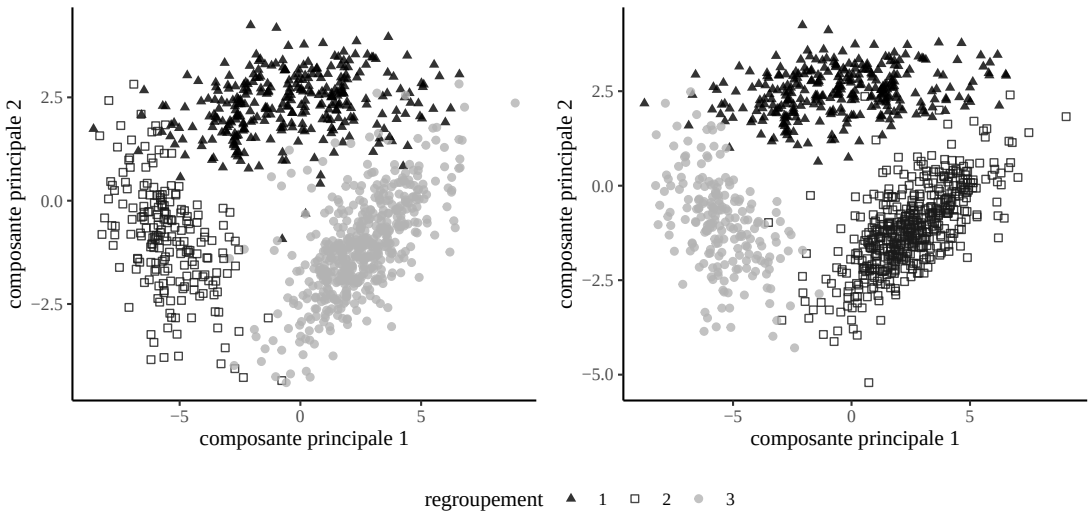

Figure 1: Projection sur deux composantes principales des regroupements obtenus par la méthode hiérarchique avec la méthode des k-moyennes avec $k = 3$ groupes. Chaque graphique représente une base de données imputées (parmi 100). Le numéro de groupe est indiqué dans la légende.

Question 4. Sélection de variables **8**

Pour chacun des énoncés suivants, indiquez s'il est **vrai ou faux**. **Justifiez** brièvement votre réponse.

4.1 Le modèle avec le plus petit BIC est toujours **meilleur** ou équivalent au modèle qui a la plus petite valeur d'AIC. [2]

> **Solution:** Faux, la pénalité étant plus élevée, le modèle retourné par le critère d'information bayésien de Schwartz est toujours plus simple, mais pas forcément meilleur pour la prédiction. Pour la sélection de modèles, la procédure est convergente.

4.2 L'avantage de la validation externe par rapport à la validation croisée est l'absence d'aléatoire: on obtient toujours la même mesure de performance ou d'erreur. [2]

> **Solution:** Faux, la séparation en bases de données d'entraînement/de validation/de test est également aléatoire.

4.3 Dans le cadre de la sélection de variables avec un ensemble de modèles linéaires candidats, on choisit celui qui a la plus petite erreur quadratique moyenne d'entraînement. [2]

> **Solution:** Faux, cette dernière est minimisée par le modèle le plus complexe de l'ensemble s'ils sont emboîtés.

4.4 La procédure LASSO mène implicitement à une sélection de variables. [2]

> **Solution:** Vrai, la forme de la pénalité $\lambda \sum_{j=1}^{p} |\beta_j|$ force certaines estimations à zéro. À mesure que la pénalité λ augmente, le nombre de variables actives (celles dont le coeffi-

cient est non-nul) diminue.

Question 5. Risque de défaillance et score de crédit **12**

Le terme de score de crédit fait référence à l'utilisation de méthodes statistiques pour classifier des créditeurs en bon ou mauvais risques. Une banque décide de faire une analyse de ses dossiers de crédits (aux particuliers) afin de mieux comprendre les caractéristiques qui font qu'une personne à qui on a accordé un prêt remboursera ce dernier avant l'échéance (`defaut=0`) ou pas (`defaut=1`).

On dispose des variables explicatives suivantes:

- `score`: score de crédit, une valeur entre 300 et 850; un score plus élevé dénote un meilleur dossier de crédit.
- `proprio`: est-ce que la personne qui a un emprunt est propriétaire, oui (`1`) ou non (`0`).
- `stabilite`: variable catégorielle pour la stabilité de l'emploi, soit faible (`faible`), modérée (`modere`) ou élevée (`eleve`).
- `tauxendet`: taux d'endettement (en pourcentage)

On divise la base de données en deux (échantillons d'entraînement et de validation) pour évaluer de manière fiable la performance.

Les coefficients du modèle logistique ajusté, les tests de significativité globale des coefficients et le tableau du lift sont présentés dans l'**Annexe 4**.

5.1 Interprétez l'effet du score de crédit sur le risque de défaut. [2]

Solution: Pour chaque augmentation du score de 1 point, la cote pour le risque de défaut/remboursement à échéance diminue de 0.54%, *ceteris paribus*. Cela représente une diminution de 41.73% par tranche de 100 points du score de crédit.

5.2 Rapportez le rapport de cote pour locataire versus propriétaire. [2]

Solution: $1/0.4636 = 2.157$

5.3 Est-ce que l'effet de la stabilité de l'emploi impacte globalement le risque de défaillance? Justifiez votre réponse. [2]

Solution: Oui. L'hypothèse nulle est que les deux coefficients associés à `stabilite` sont nuls, l'alternative qu'au moins un des deux est différent de zéro. La statistique du rapport de vraisemblance de 9.23 (à comparer à une loi khi-deux avec 2 ddl) donne une valeur-p de 1%. On rejette l'hypothèse nulle à niveau 5% pour conclure que la variable a un effet significatif.

5.4 Le lift pour le modèle de prédiction est retourné dans l'**Annexe 4**. Interprétez le lift pour 30%. [2]

Solution: Si on classe les probabilités prédites et qu'on assigne 30% des probabilités les plus élevés en succès, on détecte 1.67 fois plus qu'au hasard, une augmentation de 67%.

5.5 Votre collègue a décidé de faire de la sélection de variable et obtient trois modèles. Il calcule un point de coupure optimal pour chaque modèle. Les tableaux qui suivent résument la performance des trois modèles sur l'échantillon de validation contenant $n = 500$ observations (pour rappel, une valeur de $Y = 1$ indique un **défaut** de paiement).

Tableau 1: Tableaux de classification pour les différents modèles proposés.

	$Y=1$	$Y=0$
$\widehat{Y}=1$	2	8
$\widehat{Y}=0$	38	452

(a) Modèle 1

	$Y=1$	$Y=0$
$\widehat{Y}=1$	16	93
$\widehat{Y}=0$	24	367

(b) Modèle 2

	$Y=1$	$Y=0$
$\widehat{Y}=1$	30	40
$\widehat{Y}=0$	10	420

(c) Modèle 3

i. Si vous cherchez à minimiser le taux de mauvaise classification, quel modèle est préférable? Justifiez votre réponse. [1]

> **Solution:** Il suffit de faire le décompte de la diagonale, soit 454/500 pour Modèle 1, 383/500 pour Modèle 2 et 450/500 pour Modèle 3. Le Modèle 1 est préférable si on veut minimiser le taux de mauvaise classification

ii. La banque ne prêtera pas aux personnes qui sont identifiées comme mauvais créditeurs. Selon leur estimation, un prêt moyen permet un retour sur investissement de 2%, mais on perd en moyenne 10% si une personne ne le rembourse pas à échéance (autrement dit, il est cinq fois plus grave de classifier une observation comme bon créditeur alors que la personne fera défaut). Selon ce scénario, quel modèle est alors préférable? Justifiez votre réponse. [3]

> **Solution:** Si on considère la paire (pred, obs), on utilise comme pénalité 1 pour `00`, -5 pour `01` et zéro sinon (autres réponses possibles). Pour le modèle 1, le gain est $452 - 5 \times 38 = 262$, contre $367 - 5 \times 24 = 247$ pour le modèle 2 et $420 - 10 \times 5 = 370$ pour le modèle 3. Ce dernier (Modèle 3) est donc celui qui est préférable.