Analyse multidimensionnelle appliquée

Denis Larocque, Léo Belzile

2020-01-10

Contents

Chapter 1

Introduction

1.1 Survol du cours

1.1.1 Analyse factorielle exploratoire

On dispose de p variables $X_1, \ldots, X_p$. Peut-on expliquer les interrelations (la structure de corrélation) entre ces variables à l'aide d'un certain nombre (moins de p) de facteurs latents (non observés)?

L'analyse factorielle est souvent utilisée pour analyser des questionnaires (construction d'échelles) comme dans l'exemple suivant.

Example 1.1. Pour les besoins d'une enquête, on a demandé à 200 consommateurs adultes de répondre aux questions suivantes par rapport à un certain type de magasin:

Sur une échelle de 1 à 5,

1. pas important
2. peu important
3. moyennement important
4. assez important
5. très important

Pour vous, à quel point est-ce important…

1. que le magasin offre de bons prix tous les jours?
2. que le magasin accepte les cartes de crédit majeures (Visa, Mastercard)?
3. que le magasin offre des produits de qualité?
4. que les vendeurs connaissent bien les produits?
5. qu'il y ait des ventes spéciales régulièrement?
6. que les marques connues soient disponibles?
7. que le magasin ait sa propre carte de crédit?

8. que le service soit rapide?
9. qu'il y ait une vaste sélection de produits?
10. que le magasin accepte le paiement par carte de débit?
11. que le personnel soit courtois?
12. que le magasin ait en stock les produits annoncés?

Pouvons-nous identifier un nombre restreint de facteurs (concepts, dimensions) qui pourraient bien rendre compte de la structure de corrélation entre ces 12 variables?

Buts:

- Décrire et comprendre la structure de corrélation d'un ensemble de variables à l'aide d'un nombre restreint de concepts (appelés facteurs).
- Réduire le nombre de variables en créant une nouvelle variable par facteur. Ces nouvelles variables pourront par la suite être utilisées dans d'autres analyses (régression linéaire multiple par exemple).

1.1.2 Sélection de variables et de modèles

Dans plusieurs situations, on doit développer un modèle de prévision. Par exemple, on pourrait devoir développer un modèle pour:

- Détecter les faillites des clients (ou des entreprises)
- Cibler les clients qui seront intéressés par une offre promotionnelle
- Détecter les fraudes (par carte de crédit ou dans les rapports de revenus)
- Prévoir si un client va nous quitter.

Il y a en général plusieurs variables explicatives potentielles, et aussi plusieurs types de modèles possibles (régression linéaire, réseaux de neurones, arbres de régression ou de classification, etc.). Dans ce chapitre, nous verrons des principes généraux et des outils afin de sélectionner des modèles performants, ou bien un sous-ensemble de variables avec un bon pouvoir prévisionnel.

1.1.3 Régression logistique

On cherche à expliquer le comportement d'une variable binaire Y $(0-1)$, à l'aide de p variables quelconques $X_1, \ldots, X_p$.

Example 1.2. Une banque offre aux gens la possibilité de faire une demande de carte de crédit en ligne en promettant une approbation (conditionnelle) en quelques minutes seulement. Le tout est basé sur un modèle automatique de classification qui décide d'accorder ou non la carte ($Y = 1$ ou $Y = 0$) en fonction des réponses fournies par les clients potentiels à différentes questions comme: quel est votre revenu annuel brut (X_1), avez-vous d'autres cartes de crédit (X_2), êtes-vous locataire ou propriétaire (X_3), etc…

Buts:

- Comprendre comment et dans quelle mesure les variables X influencent la catégorie d'appartenance de Y.
- Développer un modèle pour faire de la classification, c'est-à-dire, prévoir la catégorie d'appartenance de Y pour un nouveau sujet à partir des variables X.

1.1.4 Analyse de regroupements

On cherche à créer des groupes (« *clusters* ») d'individus homogènes en utilisant p variables $X_1, \ldots, X_p$.

Example 1.3. Cette méthode est utilisée en marketing pour la **segmentation de marché**, qui consiste en

> …définir des sous-groupes réunissant des consommateurs qui partagent les mêmes préférences ou qui réagissent de façon semblable à des variables de marketing[1]

But:

- Combiner des sujets en groupes (interprétables) de telle sorte que les individus d'un même groupe soient les plus semblables possible par rapport à certaines caractéristiques et que les groupes soient les plus différents possible.

1.1.5 Analyse de survie

On s'intéresse au temps avant qu'un événement survienne. Par exemple :

- Temps qu'un client demeure abonné à un service offert par notre compagnie.
- Temps de survie d'un individu après avoir été diagnostiqué avec un certain type de cancer.
- Temps qu'un employé demeure au service de la compagnie.
- Temps qu'une franchise demeure en activité.
- Temps avant la faillite d'une entreprise (ou d'un particulier).
- Temps avant le prochain achat d'un client.

On observe chaque sujet jusqu'à ce que l'une des deux choses suivantes se produise: l'événement survient avant la fin de la période d'observation ou bien l'étude se termine et l'événement n'est toujours pas survenu. Dans le premier exemple, l'événement correspond au fait d'interrompre son abonnement. On dispose donc d'une variable temps T pour chaque individu qui est soit censurée, soit non censurée. Si l'individu a expérimenté l'événement avant la fin de la période d'observation, la valeur de est non censurée. Si l'événement n'est toujours pas survenu à la fin de la période d'observation, la valeur de T est censurée. Pour

[1] d'Astous, A. (2000). *Le projet de recherche en marketing*, 2e édition. Chenelière/McGraw-Hill.

chaque individu, on dispose également d'un ensemble de variables explicatives $X_1, \ldots, X_p$.

But:

- Étudier les effets des variables explicatives sur le temps de survie et obtenir des prévisions du temps de survie.

1.1.6 Données manquantes

Il arrive fréquemment d'avoir des valeurs manquantes dans notre échantillon.

Simplement ignorer les sujets avec des valeurs manquantes et faire l'analyse avec les autres sujets conduit généralement à des estimations biaisées et à de l'inférence invalide.

Dans ce chapitre, nous verrons une méthode très générale afin de traiter les données manquantes, l'imputation multiple. Nous verrons comment elle peut être utilisée dans un contexte d'inférence et dans un contexte de prévision.