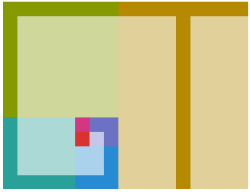

Maths – Bio

Introduction à l'apprentissage

Séance :

Alexis Jacq

Notes :

François Bienvenu

Séance du 4 décembre 2013

1 Introduction

L'utilisation de modèles est primordiale en science. Elle peut être motivée par deux objectifs différents : réaliser des prévisions, ou tester les conséquences d'une hypothèse. Mais dans ces deux cas, il existe généralement des paramètres dont les valeurs ne sont pas connues a priori et qu'il est donc intéressant d'ajuster de manière à ce que le modèle décrive plus finement possible les observations.

C'est cette questions de l'ajustement (ou apprentissage) de paramètres qui nous intéresse ici. Le but de ces séances est de présenter un cadre aussi général que possible permettant de traiter ce genre de problèmes, tout en illustrant cela par des exemples concrets. Enfin, dans la deuxième séance (cf notes séance 2), nous étudierons en détails un type de modèle particulier : les modèles à chaîne de Markov cachées.

2 Apprentissage de paramètres

2.1 Cadre général

Imaginons que l'on veuille modéliser un phénomène quelconque. La première étape sera de l'observer afin de receuillir de l'information sous la forme :

$$\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \longmapsto \begin{pmatrix} y_{11}^{\text{obs}} & \cdots & y_{1q}^{\text{obs}} \\ y_{21}^{\text{obs}} & \cdots & y_{2q}^{\text{obs}} \\ \vdots & & \vdots \\ y_{n1}^{\text{obs}} & \cdots & y_{nq}^{\text{obs}} \end{pmatrix},$$

où les entrées de la matrice $X = (x_{ij})$ correspondent à n valeurs de p variables connues en fonction desquelles on souhaite exprimer les q variables Y_j d'intérêt dont les n observations forment les colonnes de la matrice $Y_{obs} = (y_{ij}^{\text{obs}})$. Autrement dit, on observe de façon répétée (n fois) comment p variables en influencent q autres. Bien entendu, ce formalisme se veut aussi général que possible et dans la pratique la situation est souvent plus simple. Notamment, on a fréquement $q = 1$ (i.e. on n'étudie qu'une réponse à la fois) voire même $p = 1$ (i.e. on n'étudie la réponse qu'à une variable). Par exemple, on peut avoir $X = (t_i)$, représentant une variable temporelle. On parle alors de *grille de temps* pour X et de *série*

temporelle pour Y_{obs}.

Ici, notre compréhension du phénomène étudié passera par la connaissance d'un modèle $\mathcal{M}$ tel que $\mathcal{M}(X) \mapsto Y$, avec Y aussi proche que possible de Y_{obs}. Deux questions se posent alors :

- Qu'entend on par "proche" ?
- Comment trouver $\mathcal{M}$?

Bien entendu, la réponse à la première question conditionnera la réponse apportée à la deuxième. Mais puisque nous cherchons à définir un cadre aussi général que possible, nous allons nous affranchir de toute dépendance à un type de réponse particulière à la première question en supposant connue la notion de proximité. Cela passe par la connaissance d'une fonction prox telle que $\text{prox}(x, y)$ soit élevée lorsque nous considérons que x et y sont "proches", et faible lorsque nous les considérons "éloignés". Nous verrons plus tard comment les notions mathématiques de distance ou de vraisemblance peuvent être utilisées pour définir une telle fonction.

Pour répondre à la deuxième question, nous allons devoir faire une hypothèse très forte et

présupposer une certaine forme pour le modèle. Ce choix pourra être influencé soit par des hypothèses de simplicité, soit par une connaissance de certains mécanismes à la base du phénomène étudié. Nous pourrons donc être amenés à faire l'hypothèse d'un modèle linéaire ou non, ou même déterministe ou non, en fonction du phénomène particulier étudié. Quoi qu'il en soit, toujours pour rester le plus général possible, notre modèle peut s'écrire :

$$\mathcal{M}_{\mathbf{P}} : X \longmapsto Y,$$

où $\mathbf{p}$ est un vecteur de *paramètres* de notre modèle. Ce sont ces paramètres que nous allons faire varier pour trouver le "meilleur modèle" au sens de la notion de proximité que nous avons choisie, c'est à dire :

$$\mathcal{M}_{\mathbf{p}^*}\ , \quad \text{où } \mathbf{p}^* = \underset{\mathbf{p}}{\operatorname{argmax}} \left[\operatorname{prox} \left(\mathcal{M}_{\mathbf{p}}(X),\, Y_{obs} \right) \right].$$

Nous allons maintenant étudier plusieurs cas particuliers afin d'illustrer les différentes techniques qui peuvent être utilisées pour optimiser $\mathbf{p}$ en fonction du type de modèle choisi. Mais avant de poursuivre, voici un bref résumé du cadre de travail que nous avons établi :

- *Nous connaissons :* Y_{obs}, variable observée que nous cherchons à expliquer à partir de la variable X.

- *Nous avons choisi :* $\mathcal{M}_{\mathbf{p}}$, un modèle supposé adapté au problème et dont les seules inconnues sont les paramètres $\mathbf{p}$.

- *Nous cherchons :* $\mathbf{p}^*$, les paramètres pour lesquels le modèle décrit le mieux Y_{obs}, c'est-à-dire qui optimise une certaine fonction qui reste à préciser en fonction du type de modèle choisi.

2.2 Modèles déterministes

"Modèle déterministe" signifie que $\mathcal{M}_{\mathbf{p}}$ est une fonction, c'est à dire que notre *modèle* est déterministe... Mais pas forcément que le *phénomène* étudié l'est ! En effet, étant donnés X et Y_{obs}, il est toujours possible de construire un modèle décrivant *exactement* les observations, c'est-à-dire tel que $\mathcal{M}_{\mathbf{p}}(X) = Y_{obs}$. Par exemple, étant donnés n points du plan (i.e. deux vecteurs colonne X et Y_{obs}), l'utilisation des polynômes de Lagrange permet de construire un polynôme passant par chacun de ces points... Mais le polynôme résultant est de degré $n - 1$. Outre le fait que le comportement de ce modèle sera donc très probablement aberrant dès qu'on s'éloignera de l'intervalle $[\min(X), \max(X)]$,

ce polynôme contient exactement autant d'information que le couple (X, Y_{obs}) et ne nous apprend donc rien. On aurait aussi bien pu relier les points entre eux.

Plus généralement, la morale de cet exemple est qu'il ne faut pas forcément chercher le modèle décrivant le mieux les données, mais prendre aussi en compte la simplicité du modèle final. Il y a donc un compromis entre précision et simplicité, et il n'existe pas de recette générale pour le localiser.

A priori, notre modèle ne décrira donc pas parfaitement les observation. Si le phénomène observé est réellement déterministe, cela signifie que tous les mécanismes à l'oeuvre n'auront pas été pris en compte. En fonction de l'écart entre les prédiction de notre modèle ($\mathcal{M}_{\mathbf{p}}(X)$) et les observations ($Y_{obs}$), nous les aurons plus ou moins bien pris en compte. Mais le fait de permettre cet écart entre les observations et le modèle permet aussi d'utiliser un modèle déterministe pour décrire un phénomène "stochastique" (ceci est d'autant plus vrai que la stochasticité sert la plupart du temps à décrire des mécanismes déterministes que nous n'avons pas les moyens de décrire...).

2.2.1 Notion de distance

Les modèles déterministes ont en commun le fait qu'il est possible d'utiliser la notion mathématique de *distance* pour définir la proximité entre un modèle et la réalité. Une distance est une fonction vérifiant les propriétés suivantes, qui correspondent bien à l'idée intuitive que l'on se fait de la notion de distance :

- $d(x, y) = 0 \iff x = y$
- $d(x, y) = d(y, x)$
- $d(x, y) + d(y, z) \geq d(x, z)$

Par exemple, il peut être vérifié que, lorsqu'on a affaire à des vecteurs, la distance euclidienne – ou plus généralement la distance définie par :

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_i |x_i - y_i|^p \right)^{1/p},$$

où $p \in \mathbb{N}^* \cup \{+\infty\}$, vérifie bien ces propriétés. Pour une discussion plus approfondie de la notion de distance, se référer à la séance de Guilhem Doulcier sur l'analyse en groupement (clustering).

Ainsi, étant donnée une distance adaptée aux type de données considérées, on peut écrire que :

$$\mathbf{p}^* = \underset{\mathbf{p}}{\operatorname{argmin}} \left[d\left(\mathcal{M}_{\mathbf{p}}(X),\, Y_{obs}\right) \right].$$

$\mathbf{p}$ étant la seule inconnue de cette équation, on s'est ramenés à un simple problème d'optimisation. Néanmoins, la solution de ce problème sera plus ou moins simple à trouver en fonction du type de modèle et de la distance choisis. Dans la suite, nous allons détailler un cas simple et présenter rapidement une méthode permettant de traiter des cas plus complexes.

2.2.2 Modèles linéaires

Le modèle linéaire s'écrit :

$$\mathcal{M}_{\mathbf{p}}(X) = X\mathbf{p}$$

A noter que les dimensions de X, Y et $\mathbf{p}$ peuvent être quelconques, du moment qu'elles sont compatibles au niveau du produit matriciel ($X\mathbf{p} = Y$). Néanmoins, $\mathbf{p}$ et Y sont le plus fréquemment des vecteurs, ce que nous supposons dans la suite par soucis de simplicité.

Sous cette forme, le modèle n'inclut pas "d'ordonnée à l'origine" : $\mathcal{M}_{\mathbf{p}}(0_{np}) = 0_{nq}$. Néanmoins, une

astuce permet de contourner ce problème : il suffit d'ajouter une colonne ne contenant que des 1 à X et d'ajouter une entrée p_0 à $\mathbf{p}$. Ainsi, on aura bien :

$$\forall i, \quad y_i = p_0 + p_1 x_{i1} + \ldots + p_n x_{ip}.$$

Pour garder les notations simples, on note toujours X (resp. $\mathbf{p}$) la nouvelle matrice (resp. vecteur) ainsi définie.

Si l'on choisi comme distance la distance euclidienne, définie par

$$d(x, y) = \sqrt{\|x - y\|_2},$$

on obtient :

$$\mathbf{p}^* = \underset{\mathbf{p}}{\operatorname{argmin}} \sqrt{\|X\mathbf{p} - Y_{obs}\|_2}$$

Ce qui, puisque la racine est croissante, est aussi :

$$\mathbf{p}^* = \underset{\mathbf{p}}{\operatorname{argmin}} \|X\mathbf{p} - Y_{obs}\|_2$$

On va donc chercher à minimiser la somme des carrés des écarts entre le modèle et les observations – c'est ce qui vaut à cette méthode le nom de *d'estimation par les moindres carrés*, que nous appliquons ici au

cas de la *régression linéaire multiple*.

La démarche pour trouver $\mathbf{p}^*$ est simple : on commence par exprimer $\|X\mathbf{p} - Y_{obs}\|_2 = f(\mathbf{p})$ puis on cherche $\mathbf{p}^*$ tel que f ne varie pas lorsque $\mathbf{p}$ varie ($\nabla_{\mathbf{p}^*} f = (0)$, l'équivalent de "$\frac{\mathrm{d}f}{\mathrm{d}\mathbf{p}}(\mathbf{p}^*) = 0$" lorsqu'on dérive par rapport à un vecteur). Il reste ensuite à vérifier qu'il s'agit bien d'un minimum.

Calcul de $f(\mathbf{p}) = \|X\mathbf{p} - Y_{obs}\|_2$:

$$\begin{aligned} f(\mathbf{p}) &= (X\mathbf{p} - Y_{obs})^\top \ (X\mathbf{p} - Y_{obs}) \\ &= (X\mathbf{p})^\top (X\mathbf{p}) - (X\mathbf{p})^\top Y_{obs} - Y_{obs}^\top (X\mathbf{p}) + Y_{obs}^\top Y_{obs} \end{aligned}$$

Mais puisque Y_{obs} et $X\mathbf{p}$ sont des vecteurs colonnes, $(X\mathbf{p})^\top Y_{obs}$ et $Y_{obs}^\top (X\mathbf{p})$ sont des scalaires et sont donc égaux. D'où :

$$f(\mathbf{p}) = \mathbf{p}^\top X^\top X\mathbf{p} - 2\mathbf{p}^\top X^\top Y_{obs} + Y_{obs}^\top Y_{obs}$$

Dérivation : Ceux qui ont des scrupules à faire un peu n'importe quoi et à dériver par rapport à des vecteurs comme si c'était des scalaires ont raison, *même si* dans les faits on retombe souvent sur le résultat qu'on aurait obtenu en faisant un peu n'importe quoi. Il faut néanmoins garder à l'esprit que les

matrices ne commutent pas et qu'on travaille avec des dérivées partielles et non des dérivées simples, et rester vigilant. Ici, on pourra vérifier que :

$$\nabla_{\mathbf{p}} f = 2X^\top X\mathbf{p} - 2X^\top Y_{obs}$$

Remarque : Guillaume J. nous a très justement fait remarquer que nous nous sommes donné beaucoup de mal pour arriver à ce résultat pourtant très simple à montrer. En effet, on peut utiliser une astuce de calcul :

$$\text{“ } \langle u, v\rangle' = \langle u', v\rangle + \langle u, v'\rangle \text{ ”}$$

Une fois qu'on sait cela, les calculs se trouvent grandement simplifiés...

Conclusion : Lorsque le gradient d'une fonction s'annule, on parle de *point critique*. Cela ne correspond pas forcément à un optimum (il peut très bien s'agir d'un point-selle) et pour s'en assurer il est nécessaire d'étudier la hessienne ($\mathrm{H}(f)$) de la fonction (cf cours d'Amaury Lambert). Ici, on se dispense de cette étude en admettant qu'il s'agit d'un minimum, ce qu'on justifie en disant qu'il est “clair” que f est convexe. D'ailleurs, j'ai pris trente fonctions et

c'était bien le cas.

Ainsi, le gradient s'annule ssi $X^\top X\mathbf{p} = M^\top Y_{obs}$, soit, en admettant que $X^\top X$ est inversible (ce qui est le cas si les vecteurs colonnes X_i forment une famille libre) :

$$\mathbf{p}^* = (X^\top X)^{-1} X^\top Y_{obs}$$

On a donc bien trouvé une expression analytique de $\mathbf{p}^*$.

Remarque : La matrice $X^+ = (X^\top X)^{-1} X^\top$ (notée X^+ par à peu près tout le monde, à l'exception de Michael Jordan qui la note $X^\dagger$) est appelée *pseudoinverse* de la matrice X. En effet, puisqu'on cherchait, idéalement, $\mathbf{p}^*$ tel quel $X\mathbf{p}^* = Y_{obs}$, on pourrait vouloir écrire $\mathbf{p}^* = X^{-1}\, Y_{obs}$ – ce qui n'est néanmoins pas possible puisque X, n'étant pas carrée, n'est pas inversible. Mais ici, on s'est tout de même ramenés à une écriture très similaire : $\mathbf{p}^* = X^+ Y_{obs}$. De plus, il peut être vérifié que lorsque A est inversible, $A^+ = A^{-1}$. Il s'agit donc en quelque sorte d'une généralisation de la notion d'inverse.

2.2.3 Généralisation du modèle linéaire

Pour l'instant, nous nous sommes intéressé à des modèles linéaires "classiques" :

$$y_i = (p_0 +)\ p_1 x_{i1} + \ldots + p_n x_{ip}.$$

Mais, en reprenant les calculs précédents, on peut voir que c'est la linéarité par rapport à $\mathbf{p}$ qui importe et que si l'on avait eu :

$$f(y_i) = p_0 + p_1\phi_1(x_{i1}) + \ldots + p_n\phi_p(x_{ip}).$$

Les calculs seraient restés tout aussi valables et on aurait eu :

$$\mathbf{p}^* = M^+ f(Y_{obs}), \quad \text{avec } M = \begin{pmatrix} 1 & \phi_1(x_{11}) & \cdots & \phi_p(x_{1p}) \\ 1 & \phi_1(x_{21}) & \cdots & \phi_p(x_{2p}) \\ \vdots & \vdots & & \vdots \\ 1 & \phi_1(x_{n1}) & \cdots & \phi_p(x_{np}) \end{pmatrix}$$

La méthode est donc extrèmement générale. Notamment, elle peut-être utilisée pour réaliser un ajustement polynomial. En effet, la famille $(X^0, \ldots, X^{p-1})$ est libre. Par conséquent, en considérant M la matrice dont les X^k forment les colonnes, $M^\top M$ est inversible et il est donc possible de calculer M^+. D'où la possibilité d'ajustement d'un polynôme de degré souhaité.

2.2.4 Modèles non-linéaires

Il existe de nombreux cas où l'utilisation de modèles linéaires n'est pas suffisante. On peut alors avoir recours à des modèles non-linéaires, mais le problème vient alors du fait qu'il n'est pas toujours possible de résoudre $\nabla_{\mathbf{p}} f = (0)$ comme nous l'avons fait précédemment. Mais même lorsque cela n'est pas possible analytiquement, il existe diverses méthodes numériques permettant de trouver les optimums d'une fonction.

Une de ces méthodes est la *descente de gradient*. Le principe de cette méthode est simple : partant d'un (ou plusieurs) points initiaux, on va se déplacer à chaque fois dans la direction qui minimise le plus f. Plus formellement :

$$\mathbf{p}_{t+1} = \mathbf{p}_t - \alpha \nabla_{\mathbf{p}_t} f.$$

Dans cette expression, $-\nabla_{\mathbf{p}_t} f$ donne la direction de déplacement optimale au temps t, et α est un paramètre permettant de moduler la taille des déplacements. L'algorithme se termine lorsqu'une condition d'arrêt est vérifiée (typiquement, $\|\nabla_{\mathbf{p}_t} f\| \leq \varepsilon$ ou encore $t > T$).

Un exemple : Voici un exemple illustrant le principe de cette méthode : imaginons que $f : \mathbb{R}^2 \to \mathbb{R}$ représente l'altitude d'un terrain en fonction de ses coordonnées (x, y) sur une carte (par exemple, ses coordonnées dans la projection Lambert, puisque nous sommes FRANÇAIS – sinon on pourrait prendre les coordonnées GPS. D'ailleurs, si vous jetez un coup d'oeil au paragraphe "Résolution de l'équation de navigation" de l'article Wikipédia sur le GPS, vous pourrez voir qu'on voit des trucs utiles au GT !).

On cherche le point (x, y) d'altitude minimum. Pour ça, on va placer une bille qu'on laisse rouler jusqu'à ce qu'on la considère immobile ou qu'on en ait marre d'attendre. Le gradient de f est perpendiculaire à ses lignes de niveau sur une carte topographiques ($\mathcal{C}_{(x_t,y_t)} = \{(x, y) \mid f(x, y) = f(x_t, y_t)\}$) et indique donc la direction (dans $\mathbb{R}^2$) dans laquelle la pente est la plus forte (dans le sens de la montée – d'où le signe moins dans la formule de la descente de gradient). C'est donc la direction dans laquelle la bille roulerait si elle n'avait pas d'inertie (par exemple, si elle roulait infiniment lentement). On approxime donc le trajet de la bille en la faisant aller dans cette direction pour une distance qui est fonction de la pente à l'endroit où se trouve la bille, $\|\nabla_{(x_t,y_t)} f\|$,

et du paramètre α qui peut être interprété comme la "vitesse" de la bille du moment qu'on ne pousse pas l'analogie trop loin.

Ainsi, si α n'est pas trop grand et si le terrain n'est pas trop irrégulier, on reproduira bien le trajet d'une bille roulant infiniment lentement sur ce terrain et terminant sa course au niveau d'un minimum. Néanmoins, si le terrain est trop irrégulier (par exemple, contient des vallées très étroites) et qu'α n'est pas assez petit, on pourra passer au travers d'une vallée sans y rester – un peu comme le ferait une bille ayant une vitesse élevée qui roulerait dans la vallée mais remonterait de l'autre côté du fait de son inertie.

On voit bien que le principal problème de cette méthode est lié au fait qu'on est jamais sûr d'avoir atteint *le* minimum absolu : on ne fait que trouver des minimums locaux. Par exemple, même si vous lancez votre bille à proximité du grand canyon, rien ne vous indique qu'elle arrivera en bas car elle peut très bien rester bloquée dans une petite crevasse située sur le plateau. Pour éviter cela, une solution est de partir de plusieurs endroits, idéalement "le plus que possible avec α aussi petit que possible", pour se don-

ner une chance de trouver les minimums ayant des bassins d'attraction très réduit (ex : un trou sur un parcours de golf).

En fait, il existe de nombreuses méthodes pour échantillonner l'espace de façon efficace, et la méthode du gradient n'est pas la seule méthode pouvant être utilisée pour la recherche de minimums, mais cela dépasse le cadre de ces séances du groupe de travail.

2.3 Modèles probabilistes

Pour l'instant, nous n'avons considéré que des modèles déterministes. Mais il existe des cas où il est plus intéressant de prendre en compte la stochasticité de façon explicite dans le modèle – soit parce que cela permet un meilleur ajustement du modèle (via un meilleur traitement des valeurs extrêmes par exemple), soit parce que les paramètres régissant la stochasticités sont des paramètres intéressants nous renseignant sur le phénomène étudié.

En rédigeant ces notes, je me suis rendu compte du fait qu'autant il semble très simple de distinguer un modèle déterministe d'un modèle stochas-

tique, autant il n'est pas facile de mettre le doigt sur les différences fondamentales entre ces types de modèles : nous avons vu que les modèles déterministes sont souvent utilisés pour modéliser des phénomènes non déterministes (dans l'exemple de la régression linéaire, l'écart des observations à la droite de régression est souvent dû à du "bruit" (erreurs de mesures, etc) de nature stochastique. Dans quelle situation est-il alors vraiment important d'inclure la stochasticité dans le modèle ? Enfin, les modèles déterministes peuvent être envisagés comme des cas particuliers de modèles stochastiques. Quelles sont les hypothèses faites par rapport au modèle stochastiques lorsqu'on travaille avec le modèle déterministe correspondant ? De même, quand et comment peut-on se ramener à un modèle déterministe (par exemple, choisir de minimiser une certaine distance peut en fait revenir à faire une hypothèse sur la stochasticité du système). Si vous ne voyez pas ce que je veux dire, considérez l'exemple de la régression linéaire. Nous l'avons considérée comme un modèle déterministe, et avons posé *a priori* qu'une le meilleur modèle serait celui minimisant la distance euclidienne entre les observations et le modèle. En fait, cela revient à considérer un modèle probabiliste en faisant un certain

nombre d'hypothèses (écarts au modèle d'espérance nulle, non corrélés et de même variance : cf théorème de Gauss-Markov ou encore la section "hypothèses" de l'article Wikipédia sur la régression linéaire). Le but de ce paragraphe était simplement de dire que lorsqu'on y réfléchit, la question s'avère plus intéressante qu'il n'y paraît de prime abord.

Quoi qu'il en soit, la particularité des modèles probabilistes sera qu'au lieu de postuler que le meilleur modèle est celui minimisant une certaine distance, nous allons faire certaines hypothèses sur la stochaticité du phénomène que nous utiliserons ensuite pour en déduire le modèle le plus *probable*. Comme nous venons le voir, cela pourra, en fonction des hypothèses faites, nous faire retomber sur un modèle déterministe.

2.3.1 Cadre général

Un modèle probabiliste prend en compte l'aléatoire de façon explicite en considérant que les observations $(y_{11}^{\text{obs}}, \ldots, y_{1q}^{\text{obs}}), \ldots, (y_{n1}^{\text{obs}}, \ldots, y_{nq}^{\text{obs}})$ sont les réalisations des variables aléatoires $Y_1, \ldots, Y_n$. En général, on suppose que ces variables aléatoires sont

indépendantes et identiquement distribuées (iid). Il s'agit d'une hypothèse très forte mais généralement justifiée (sauf dans certains cas comme les séries temporelles où, si le système possède une certaine mémoire, on peut s'attendre à avoir de la corrélation entre les Y_i – nous verrons des exemples de cela dans la partie sur les chaînes de Markov cachées).

Le choix du type de modèle consiste à choisir une loi pour les Y_i. Cette loi dépend de paramètres $\mathbf{p}$ que nous allons optimiser de manière à décrire le mieux possible les données.

Reste à définir la "proximité" du modèle aux données. Dans le cas déterministe, nous avions utilisé une fonction de distance. Dans le cas de modèles probabilistes, cela n'est – a priori – plus adapté car ces modèles ne réalisent pas des prédiction ponctuelles de ce que devraient être les observations mais de ce que devrait être leur distribution. Il faut donc utiliser une fonction quantifiant l'adéquation des données avec les distributions prévues par le modèle.

Une idée naturelle est d'utiliser $\mathbb{P}[Y_{obs} \mid \mathbf{p}]$, la probabilité des observations Y_{obs} sachant le modèle

(i.e. "le modèle étant ce qu'il est"). Puisque nous voulons optimiser $\mathbf{p}$, Y_{obs} étant fixé, il est plus naturel d'exprimer cette probabilité comme une fonction de $\mathbf{p}$ paramétrée par Y_{obs} :

$$\mathcal{L}(\mathbf{p} \mid Y_{obs}) = \mathbb{P}[Y_{obs} \mid \mathbf{p}].$$

Cette fonction est appelée la vraisemblance (ou *likelihood*) de $\mathbf{p}$ sachant Y_{obs}. On peut chercher à la maximiser et définir $\mathbf{p}^*$ comme :

$$\mathbf{p}^* = \underset{\mathbf{p}}{\operatorname{argmax}}\, \mathcal{L}(\mathbf{p} \mid Y_{obs}).$$

Cette méthode, très utilisée, est appelée *estimation par maximum de vraisemblance* (MLE). Comme nous allons le voir par la suite, elle présente des avantages (notamment sa facilité d'utilisation). Néanmoins, elle a aussi de nombreuses limites et on peut être amené à lui préférer d'autres méthodes comme l'inférence Bayésienne, qui vise à maximiser $\mathbb{P}[\mathbf{p}|Y_{obs}]$. La raison pour laquelle l'inférence Bayésienne n'est pas systématiquement préférée à l'estimation par maximum de vraisemblance est qu'elle est souvent beaucoup plus difficile à utiliser. En effet, autant le modèle donne directement $\mathbb{P}[Y_{obs}|\mathbf{p}]$, autant il ne donne pas $\mathbb{P}[\mathbf{p}|Y_{obs}]$, qui n'est connue que par :

$$\mathbb{P}[\mathbf{p}|Y_{obs}] = \frac{\mathbb{P}[Y_{obs} \mid \mathbf{p}]\,\mathbb{P}[\mathbf{p}]}{\mathbb{P}[Y_{obs}]},$$

le problème étant que cette formule fait intervenir $\mathbb{P}[\mathbf{p}]$ et $\mathbb{P}[Y_{obs}]$, qui ne sont pas connues dans la plupart des cas. En revanche, lorsqu'elles le sont, l'estimation Bayesienne s'avère bien plus puissante que l'estimation par maximum de vraisemblance. L'exemple classique du test de dépistage (qui a été discuté en détail lors de la séance sur les modèles matriciels de population) illustre bien cela.

Le maximum de vraisemblance souffre d'une autre limitation, qui est plus apparente lorsqu'on travaille en continu. En effet, on a alors $\mathbb{P}[Y_{obs} \mid \mathbf{p}] = 0$ et il faut travailler avec la densité $f(Y_{obs} \mid \mathbf{p})$. La méthode du maximum de vraisemblance consiste à chercher le maximum de f. Mais il se peut très bien que ce maximum absolu, atteint pour $\mathbf{p}^*$, corresponde à un pic isolé et qu'il existe par ailleurs une large bosse centrée sur $\widetilde{\mathbf{p}}$, de sorte qu'il existe l tel que pour $\varepsilon > l$, $\mathbb{P}\left(Y_{obs} \mid \mathbf{p} \in [\widetilde{\mathbf{p}} - \varepsilon, \widetilde{\mathbf{p}} + \varepsilon]\right) > \mathbb{P}\left(Y_{obs} \mid \mathbf{p} \in [\mathbf{p}^* - \varepsilon, \mathbf{p}^* + \varepsilon]\right)$ – auquel cas, selon la situation étudiée, il pourrait être plus intéressant de travailler avec $\widetilde{\mathbf{p}}$, qui donne une idée de la position de la "famille de modèles les plus probables" dans l'espace des paramètres.

Malgré ces limitations, l'estimation par maximum

de vraisemblance reste une méthode puissante. Nous allons maintenant donner deux exemples simples de son utilisation.

2.3.2 Paramètres d'une loi normale

Supposons que $\forall i,\ Y_i \sim \mathcal{N}(\mu, \sigma^2)$. Ici,

$$\mathbf{p} = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} \quad \text{et} \quad \mathbf{p}^* = \underset{\mathbf{p}}{\operatorname{argmax}}\, f(Y_{obs} \mid \mathbf{p}),$$

où f est la densité de probabilité de Y. Les Y_i étant iid, on a :

$$f(Y \mid \mathbf{p}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}}.$$

L'idée est une fois encore de dériver pour trouver le(s) optimum(s) potentiel de f et de s'assurer ensuite qu'on a affaire à un maximum. Mais avant cela, comme il n'est pas très pratique de travailler avec des produits (surtout pour dériver !), on préfère se ramener à une somme en passant au log. Ceci ne change pas les valeurs de $\mathbf{p}$ pour lesquels d'éventuels optimums seraient atteint car la fonction log est strictement croissante (et donc préserve les

inégalités). On note alors :

$$\mathfrak{L} = \sum_{i=1}^{n} -\log(\sqrt{2\pi\sigma^2}) - \frac{(y_i - \mu)^2}{2\sigma^2}$$

Il s'agit de l'expression de la log-vraisemblance de notre modèle en fonction de μ et σ^2.

$$\frac{\partial \mathfrak{L}}{\partial \mu} = 0 \iff \sum_{i=1}^{n} \frac{(y_i - \mu)}{\sigma^2} = 0 \iff \mu^* = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Cette valeur est indépendante de σ^2 et on reconnait l'expression de la moyenne empirique, qui est bien l'estimateur usuel de l'espérance. On peut remarquer qu'on est sûr d'avoir affaire à un maximum dans la diretion de μ car $\forall i,\ \mu \mapsto -(y_i - \mu)^2$ est concave et donc en tant que somme de fonctions concaves, $\mu \mapsto \mathfrak{L}(\mu, \sigma^2)$ l'est également.

On cherche maintenant le maximum selon σ^2. $\mathfrak{L}$ peut être réécrit :

$$\mathfrak{L} = -\frac{n}{2}\left(\log(2\pi) + \log(\sigma^2) + \frac{v(\mu)}{\sigma^2}\right),$$

où $v(\mu) = \frac{1}{n}\sum(y_i - \mu)^2$. Ainsi,

$$\frac{\partial \mathfrak{L}}{\partial \sigma^2} = -\frac{n}{2}\left(\frac{1}{\sigma^2} - \frac{v(\mu)}{(\sigma^2)^2}\right),$$

d'où

$$\frac{\partial \mathfrak{L}}{\partial \sigma^2} = 0 \iff \frac{1}{\sigma^2} = \frac{v(\mu)}{(\sigma^2)^2} \iff \sigma^{2*} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu^*)^2$$

On retrouve là encore l'expression "intuitive" de la variance empirique (après avoir vérifié qu'il s'agit bien d'un minimum, ce qui se fait facilement en regardant le signe de $\frac{\partial \mathfrak{L}}{\partial \sigma^2}$ de part et d'autre de cette valeur.

Remarque importante Le but du paragraphe précédent était simplement d'illuster la méthode, qui ne s'avère pas très intéressante ici : en effet, le résultat concernant l'estimateur de la moyenne pouvait être retrouvé très facilement en vérifiant que la moyenne empirique $\overline{Y} = \frac{1}{n}(Y_1 + \ldots + Y_n)$ est un estimateur sans biais ($\mathbb{E}\left(\overline{Y}\right) = \mu$, immédiat par linéarité et iid des Y_i) et convergent ($V\left(\overline{Y}\right) \to 0$ quand $n \to +\infty$: en effet, $V\left(\overline{Y}\right) = \frac{1}{n^2} V\left(Y_1 + \ldots + Y_n\right)$. Or par iid, $V\left(Y_1 + \ldots + Y_n\right) = (\sigma^2 + \ldots + \sigma^2) = n\sigma^2$, d'où $V(\overline{Y}) = \frac{\sigma^2}{n}$). Quand au résultat sur la variance, il n'est pas optimal. En effet, on peut montrer que le bon estimateur de la variance est $S^2 =$

$\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y})^2$. En effet :

$$\mathbb{E}\left[\sum_{i=1}^{n}(Y_i-\overline{Y})^2\right]=\sum_{i=1}^{n}\mathbb{E}\left[Y_i^2\right]-2\,\mathbb{E}\left[Y_i\overline{Y}\right]+\mathbb{E}\left[\overline{Y}^2\right]$$

or :

$$\mathbb{E}\left[Y_i\overline{Y}\right]=\tfrac{1}{n}\,\mathbb{E}[Y_iY_1+...+Y_iY_i+...Y_iY_n]=\tfrac{n-1}{n}\,\mathbb{E}[Y]^2+\tfrac{1}{n}\,\mathbb{E}\left[Y^2\right],$$

par indépendance de Y_i et Y_j pour $i\neq j$. De même,

$$\begin{aligned}\mathbb{E}\left[\overline{Y}^2\right]&=\tfrac{1}{n^2}\,\mathbb{E}[(Y_1Y_1+...+Y_1Y_n)+...+(Y_nY_1+...Y_nY_n)]\\&=\frac{n-1}{n}\,\mathbb{E}\left[Y\right]^2+\frac{1}{n}\,\mathbb{E}\left[Y^2\right]\end{aligned}$$

d'où

$$\mathbb{E}\left[\sum_{i=1}^{n}(Y_i-\overline{Y})^2\right]=(n-1)\left(\mathbb{E}\left[Y^2\right]-\mathbb{E}\left[Y\right]^2\right)$$

Finalement,

$$\mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y})^2\right]=V(Y)$$

A noter que, de plus, ce résultat est plus général puisqu'on a seulement utilisé l'hypothèse d'iid (et

pas le fait que les Y_i suivent des lois normales). On voit donc que l'estimation par maximum de vraisemblance n'est pas toujours optimale. Néanmoins, on peut aussi remarquer que les estimateurs obtenus sont très proches des estimateurs usuels dès que n est assez grand.

2.3.3 Paramètres d'une loi multinomiale

On suppose maintenant qu'on dispose de n variables $Y_1, \ldots, Y_n$ iid pouvant prendre un nombre fini de valeurs $v_1, \ldots, v_k$ et prenant chacune de ces valeurs avec les probabilité $\pi_1, \ldots, \pi_k$, respectivement. Ce modèle est très utile car c'est le modèle le plus simple pour une variable prenant un nombre fini de valeurs. Il est utilisé en biologie dans de très nombreuses situations (modèles des taux de mutations des différents nucléotides de l'ADN, etc...).

Si l'on dispose des observations directes des variables Y_i, par indépendance, la vraisemblance des paramètres $\boldsymbol{\pi}$ s'écrit :

$$\mathcal{L}(\boldsymbol{\pi} \mid Y_{obs}) = \prod_{i=1}^{n} \mathbb{P}[Y_i = y_i^{\text{obs}} \mid \boldsymbol{\pi}] = \prod_{i=1}^{n} \pi_{\phi_i},$$

où ϕ_i est l'indice de la valeur $v_1, \ldots, v_k$ prise par y_i^{obs}.

Cela peut également s'écrire :

$$\mathcal{L}(\boldsymbol{\pi} \mid Y_{obs}) = \prod_{j=1}^{k} {\pi_j}^{c_j},$$

où c_j est le nombre de y_i^{obs} valant v_j.

Remarque : Il se peut qu'on ne dispose pas des observations 'directes' des Y_i (c'est-à-dire de la séquence $y_1^{\text{obs}}, \ldots, y_n^{\text{obs}}$) mais simplement des données de comptage de ces observations, i.e. de la variable $C = (C_1, \ldots, C_k)$, où C_j est la variable aléatoire égale au nombre de variables Y_i ayant pris la valeur v_j. Dans ce cas, on dit que C suit une loi multinomiale de paramètres $n, \pi_1, \ldots, \pi_k$ (Note : en Français, le terme "loi multinomiale" semble légèrement plus restreint. Par exemple, comparer la définition donnée sur http://fr.wikipedia.org/wiki/Loi_multinomiale et sur http://en.wikipedia.org/wiki/Multinomial_distribution).

Mais, non seulement il est également possible d'utiliser la technique du maximum de vraisemblance pour estimer les paramètres $\boldsymbol{\pi}$ dans ce cas, mais le calcul est identique à celui pour le cas où on dispose

de la séquence complète des Y_i. En effet, la vraisemblance s'écrit alors :

$$\mathcal{L}(\boldsymbol{\pi} \mid Y_{obs}) = \mathbb{P}[C_1 = c_1, \ldots, C_k = c_k] = \binom{k}{c_1, \cdots, c_k} \times \prod_{j=1}^{k} {\pi_j}^{c_j},$$

où

$$\binom{k}{c_1, \cdots, c_k} = \frac{k!}{c_1! \ldots c_k!}$$

est le coefficient multinomial de paramètres $k, c_1, \ldots, c_k$ et correspond au nombre d'ordre dans lesquels les valeurs v_k ont pu arriver. A la limite, la seule chose qui nous intéresse ici sur ce coefficient multinomial est qu'il est indépendant de $\boldsymbol{\pi}$. En effet, on ne cherche que la valeur de $\boldsymbol{\pi}$ pour laquelle un maximum est atteint, sans se soucier de la valeur de ce maximum.

Ainsi, dans les deux cas, on cherche :

$$\boldsymbol{\pi}^* = \underset{\boldsymbol{\pi}}{\operatorname{argmax}} \prod_{j=1}^{k} \pi_j^{c_j}$$

Soit, après passage au log :

$$\boldsymbol{\pi}^* = \underset{\boldsymbol{\pi}}{\operatorname{argmax}} \sum_{j=1}^{k} c_j \log(\pi_j)$$

Mais attention ! Il ne faut pas oublier qu'on doit toujours avoir $\sum_{i=1}^{k} \pi_i = 1$. On doit donc maximiser l'expression précédente *sous une contrainte*. Une façon de faire cela pourrait être d'utiliser le fait que les π_i somment à 1 pour remplacer l'un d'entre eux (par exemple π_k) par une fonction des autres ($f(\pi_1, \dots \pi_k)$). Ceci permettrait de se ramener à un problème d'optimisation sans contrainte par rapport à $(k-1)$ variables. Néanmoins, on sent bien qu'avec cette méthode les calculs risquent d'être assez lourds.

En fait, il existe une méthode assez générale pour résoudre des problèmes d'optimisation sous contrainte : la *méthode du Lagrangien* (ou *méthode des multiplicateurs de Lagrange*). La présentation détaillée de cette méthode dépasse le cadre de cette séance du GT, donc on se contentera d'une présentation "intuitive".

Digression : méthode d'optimisation du Lagrangien Le principe de la méthode du Lagrangien et de ramener un problème d'optimisation sous contrainte à un simple problème d'optimisation en ajoutant à la quantité qu'on veut optimiser des quantités qui ne seront optimisées que lorsque les con-

traintes seront vérifiées. Ainsi, plutôt que d'éliminer les contraintes en exprimant certaines variables en fonctions des autres, on les élimine *en rajoutant des variables par rapport auxquelles optimiser* – l'avantage venant du fait que les calculs résultants sont plus simples.

L'intuition derrière cette méthode est assez simple à sentir de façon graphique. Reprenons l'exemple de notre fonction $f : \mathbb{R}^2 \to \mathbb{R}$ qui aux coordonnées GPS (ou Lambert, si vous y tenez) d'un point associe son altitude. On peut se représenter cette fonction au moyen d'une carte topographique, sur laquelle sont tracées les courbes de niveau de f, $C_z = \{(x, y) \in \mathbb{R}^2 \mid f(x, y) = z\}$. La contrainte quant à elle correspond à une restriction de la recherche des optimums à un sous ensemble de $\mathbb{R}^2$. Ici, on ne considère que des contraintes "simples", i.e. consistant à restreindre la recherche à une variété (ou courbe en dimension 1, surface en dimension 2 etc – bien sûr, ici la dimension s'applique à la variété et pas à l'espace des variables : si on impose $n - 1$ contraintes en dimension n, notre variété sera une courbe, si on en impose que $n - 2$ il s'agira d'une surface, etc). Concrètement, on dispose donc d'un système d'équations $g_i(x_1, ..., x_2) = \gamma_i$. Dans notre

exemple, l'espace des variables étant de dimension 2, on ne peut avoir qu'une seule contrainte, $g(x, y) = \gamma$. Celle-ci peut être représentée par un trait sur notre carte. La méthode d'optimisation par les multiplicateurs de Lagrange repose sur le fait qu'un optimum local sous contraintes est atteint lorsque la courbe de contrainte est tangente à une courbe de niveau de f. Un dessin permet de voir ce qui se passe : l'optimum qu'on cherche se trouve forcément sur la courbe de contrainte. S'il était atteint pour un point où cette courbe n'est pas tangente à une courbe de niveau de f, alors il serait possible, en se décalant un petit peu dans le bon sens sur la courbe de contrainte, de passer au dessus (ou en dessous) de la ligne de niveau de f, auquel cas on aurait trouvé une position sur la courbe de contrainte où la valeur de f serait plus élevée (resp. plus faible), ce qui serait en contradiction avec le fait qu'on était sur un optimum. La seule façon d'empêcher le randonneur de monter ou de descendre sur tout en restant sur la courbe de contrainte, et que celle ci *ne traverse pas* de courbe de niveau à l'endroit considéré. Si tout ceci vous semble peu clair, c'est sans doute que vous n'avez pas pris le temps de faire un dessin...

Si on essaie de formaliser un peu ce qu'on vient

de voir :

- le point $\mathbf{x}^*$ pour lequel l'optimum est atteint doit vérifier $g(\mathbf{x}^*) = \gamma$
- en $\mathbf{x}^*$, la courbe de contrainte (définie par $g(\mathbf{x}) = \gamma$) doit être tangente à une courbe de niveau de f (définie par $f(x) = d$). Autrement dit, il doit exister λ tel que : $\nabla_{\mathbf{x}^*} f = \lambda \nabla_{\mathbf{x}^*} g$.

Ceci nous amène donc à introduire :

$$\Lambda(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda(g(\mathbf{x}) - \gamma)$$

Il s'agit du Lagrangien de f sous la contrainte $g - \gamma$. C'est une fonction à $n+1$ variable, où n est le nombre de variables de f. On vérifie bien que :

$$\nabla\Lambda = (0) \iff \begin{cases} \forall i, \frac{\partial f}{\partial x_i} = \lambda \frac{\partial g}{\partial x_i} & \text{(optimisation)} \\ \\ g(\mathbf{x}) - \gamma = 0 & \text{(contrainte)} \end{cases}$$

En fait, on pourrait introduire autant de contraintes que souhaité en introduisant la fonction :

$$\Lambda(x_1,...,x_n,\lambda_1,...,\lambda_k) = f(x_1,...,x_n) - \sum_{i=1}^{k} \lambda_i (g(x_1,...,x_n) - \gamma_i)$$

On peut même étendre cette méthode à des contraintes décrites par des inégalités.

Ici, on a fait que présenter cette méthode de façon rapide, sans rentrer dans les détails. En réalité, il existe des conditions sur les contraintes pour que cette méthode soit valable (dites conditions de qualification de la contrainte). Nous ne rentrons pas dans ces détails. Ceux qui le souhaitent pourront facilement trouver plein d'information à ce sujet sur Internet (hélas, il semble que la plupart des textes sur le sujet sont soit assez techniques, soit en lien direct avec des notions d'économie pas forcément très fun...)

Retour au maximum de vraisemblance On applique maintenant la méthode du Lagrangien à notre problème :

$$\Lambda(\boldsymbol{\pi}, \lambda) = \sum_j c_j \log \pi_j + \lambda \left(\sum_j \pi_j - 1 \right)$$

D'où :

$$\nabla \Lambda = (0) \iff \begin{cases} \forall i, \pi_i = \frac{c_i}{\lambda} \\ \\ \sum_j \pi_j = 1 \end{cases}$$

En substituant les π_i par leur valeur dans la contrainte, on trouve $\lambda = \sum_j c_j$, qu'on resubstitue dans

l'expression de π_i :

$$\pi_i^* = \frac{c_i}{\sum_j c_j}$$

Il resterait à vérifier qu'il s'agit bien d'un maximum, ce qu'on admet ici (de toutes façons on voit bien qu'en tant que somme de fonctions concaves, la vraisemblance est concave...)

En conclusion, on a montré ici que l'estimation par maximum de vraisemblance nous conduisait à estimer les probabilités par les fréquences. A ceux qui trouvent qu'on s'est donné beaucoup de mal pour pas grand chose, on peut répondre que le but ici était simplement d'illustrer la méthode du maximum de vraisemblance sur un exemple simple (en en profitant au passage pour voir une technique intéressante). Et à ceux qui serait très enthousiasmes et penserait qu'on a démontré que "fréquence $\sim$ probabilité", en réalité ceci ne se démontre pas : les axiomes de la théorie des probabilités sont *choisis* de manières à ce que cela soit vrai...