Rapport initiation à la recherche: l'Apprentissage Multi-Tâche

Gabriel Legout

19 décembre 2025

Résumé

Dans ce document, nous commençons par décrire et analyser des techniques d'apprentissage multi-tâche dans le cadre linéaire en se basant sur l'article ≪ Provable Meta-Learning of linear representation ≫ et en explicitant certains théorèmes et certaines techniques qui permettent de récupérer efficacement les paramètres d'estimation de notre modèle.

Table des matières

1 Introduction **3**
1.1 Conventions 4

2 Méthodes et hypothèses utilisés dans l'article **4**
2.1 La méthode des moments 4
2.2 Les hypothèses 1 et 2 5

3 Détail algorithme 1 : Calcul et borne d'un estimateur de B **5**
3.1 Algorithme 1 5
3.1.1 Un retour sur la preuve du Lemme 2 5
3.1.2 Le lemme 2 et passage à une représentation de B 8
3.1.3 De la représentation de B à l'algorithme 1 8
3.2 Éléments de preuve du théorème 3 : une première borne sur $\hat{\mathbf{B}}$ 8

1 Introduction

Commençons d'abord par réintroduire les méthodes de regression linéaire et multi-linéaire ainsi que les différents enjeux associés dans les méthodes d'apprentissages.

D'abord, de manière classique, on peut noter un modèle de regression linéaire unidimensionel de la manière suivant :

$$y_i = x_i\alpha + \epsilon_i$$

Avec, $y_i \in \mathbb{R}, x_i \in \mathbb{R}, \epsilon_i \in \mathbb{R}$. On a donc une liste de points $(x_i)_{i\in[|0,n|]}, n \in \mathbb{N}$ auquel on associe un unique α qui va être calculer selon une méthode quelconque (on peut chercher α tel qu'il minimise la somme résiduelle des carrés, auquel cas on aura y_i qui sera une prédiction au sens des moindres carrés) et va nous permettre de prédire un y_i qui n'est pas dans le domaine de points initial mais qui répond aux mêmes contraintes sur la tâche considérée.
Comme tous les modèles de regression, il faut connaître la fonction suivi par les donnés afin de pouvoir les modéliser, et en particulier, il faut que les donnés suivent une même "fonction" qui est exprimable dans le modèle.
On constate facilement qu'on peut étendre cette méthode à de la regression sur des donnés plus importantes. On peut par exemple condidérer des vecteurs à la place des scalaire et avoir un modèle multi-dimensionnel. On peut exprimer ce type de modèles sous cette forme :

$$y_i =$$

On peut encore une fois essayer de généraliser ce modèle en considérant l'hypothèse selon laquelle les données partage des caractéristiques commune qui pourrait permettre de considérer un même modèle faisant des prédictions sur des tâches différentes à partir des mêmes donnés. Dans la suite, on va considérer que chaque lien entre donnés est linéaire et que nos donnés contiennent suffisament d'information pour faire des la prédictions sur des tâches différentes. On aboutit alors au modèle suivant.

$$y_i =$$

Cependant, on voit que ce modèle aboutit à un nombre considérable de points ce qui peut entrainer au coût de calcul énorme. De plus, avec l'hypothèse selon laquelle les entrées contiennent suffisament d'information pour faire de la prédiction sur plusieurs tâches, il parait raisonable de considérer que certaines de ces tâches contiennent des caractéristiques commune parmi les entrées. Ce qu'on peut faire c'est imaginer avoir des paramètres communs à toutes les tâches et avoir des coefficients d'ajustement, par tâches, pour permettre une meilleure estimation.

Maintenant, on va donner puis expliquer dans les grandes lignes le modèle sur lequel nous travaillons :

$$y_i = \mathbf{x}_i^T \mathbf{B}\boldsymbol{\alpha}_{t(i)} + \epsilon_i$$

Dans ce modèle, on dispose d'un vecteur d'entrée comprenant d variables et de vecteurs de paramètres $\boldsymbol{\alpha}_{t(i)}$ spécifiques à une tâche i permettant de prévoir la sortie y_i.
De plus, contrairement à un modèle linéaire classique, on dispose d'une matrice $\mathbf{B}$, **commune** à toutes les tâches qui comprend des caractéristiques communes à plusieurs tâches permettant de mutualiser l'apprentissage et la prédiction y_i. Les indices $t(i)$, permettent simplement d'indicer les tâches.

Tout l'enjeu de l'article et de montrer qu'il existe des méthodes simples, sous certaines bonnes hypothèses, pour récupérer un estimateur $\hat{\mathbf{B}}$ de la matrice $\mathbf{B}$ efficacement et puis de montrer qu'on a des meilleures estimations avec un coût plus faible qu'en cherchant uniquement des $\boldsymbol{\alpha}_t$. L'article se décompose donc en deux grandes parties : d'abord la phase de "Meta-Learning" où on va essayer de trouver $\hat{\mathbf{B}}$ avec suffisament de précision et avec un algortihme simple et efficace, puis la phase de "Meta-Test" qui va permettre de montrer que la calcul de $\hat{B}$ va permettre de transférer une part de l'apprentissage à une nouvelle tâche $\boldsymbol{\alpha}_{t+1}$.

1.1 Conventions

Par soucis de clarté et contrairement à l'article, on va utiliser que $\mathbb{R}^n$ est un n-uplet de la forme $\mathbf{x} \in \mathbb{R}^n$ tel que $\mathbf{x} = (x_1, ..., x_n)$ et donc dans ce cas, $\mathbf{x}_i \in M_{d,1}(\mathbb{R})$ et donc $\mathbf{x}_i^T \in M_{1,d}(\mathbb{R})$. On a donc, avec ces conventions, $\mathbf{B} \in M_{d,r}(\mathbb{R})$ et $\boldsymbol{\alpha}_{t(i)} \in M_{r,1}(\mathbb{R})$ afin d'avoir $y_i \in \mathbb{R}$ et $\epsilon_i \in \mathbb{R}$.

On s'autorisera parfois à noter $t(i) = t$ qui restera implicitement une fonction de i afin d'alléger les notations.
Les $(\mathbf{x}_i)$ représentent des vecteurs obervations. Quand ça sera nécessaire (pour choisir un représentant pour une espérance car ils suivent tous la même loi par exemple), on s'autorisera à noter $\mathbf{x}_1 = \mathbf{x}$ un vecteur quelconque de $(\mathbf{x}_i)$ pour alléger encore une fois les notations.

Ensuite, dans tout ce document, le symbol ˆ(chapeau) au dessus d'une lettre designera un estimateur de l'objet considéré. Des théorèmes comme la loi forte des grands nombres nous assurent par exemple que la moyenne empirique est un estimateur convergent de l'espérance d'une variable aléatoire (convergence presque sûre) mais nous passerons sur le détail des convergences dans la suite de ce texte et on supposera que nous manipulerons uniquement des estimateurs convergent et on notera donc directement ces estimateurs avec un ˆ sans plus de justifications sauf quand il y aura des ambiguités.

2 Méthodes et hypothèses utilisés dans l'article

Nous allons ici exposer les hypothèses faites sur le modèles et sur les entrées ainsi que des méthodes générales qui vont être utiliser pour montrer les résultats important tout au long de l'article.

2.1 La méthode des moments

Le but de la méthode des moments est d'estimer un paramètre λ en remplaçant son moment théorique par un moment empirique. On a alors que le moment empirique de la variable aléatoire converge presque-sûrement vers l'espérance de la variable aléatoire par la loi forte des grands nombres.
Par exemple : Supposons que X suit une loi exponentielle de paramètre λ, m $= \mathbb{E}[X] = \frac{1}{\lambda}$ alors on a pour n suffisament grand :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{et} \quad \overline{X} = \hat{m}$$

D'où

$$\frac{1}{\hat{\lambda}} = \overline{X} \Leftrightarrow \hat{\lambda} = \frac{1}{\overline{X}}$$

2.2 Les hypothèses 1 et 2

L'hypothèse d'avoir les $\mathbf{x}_i$ centrés réduits est nécessaire à presque toutes les preuves pour simplifier les moments théoriques d'ordre 1 ou simplifier par l'identité les produits $\mathbf{x}\mathbf{x}^T$ dans des espérances. De plus, on considère qu'ils sont indépendants et identiquement distribués (iid).
L'hypothèse 2 permet de garantir que les tâches ne sont pas trop différentes et que donc il est possible de reconstruire la matrice $\mathbf{B}$ (qu'il y ait suffisament d'information, ou de signal si on parle de reconstruction de phase, par rapport au bruit)
L'hypothèse de sous-exponentielle permet d'appliquer des inégalités de concentration.
Pour l'utiliser, il faut alors disposer de l'expression de $\mathbb{E}$ où remplacer la valeur de l'espérance par la valeur empirique et devoir inverser l'espérance (ou en produire une estimation.)

3 Détail algorithme 1 : Calcul et borne d'un estimateur de B

On va donc décrire et expliciter ici certains points des preuves et de la méthode qui va nous permettre de calculer $\hat{\mathbf{B}}$.

3.1 Algorithme 1

L'algorithme 1 nous dit que faire une décomposition en valeurs singulières (ici la matrice est évidement carré symétrique et même définie positive donc cela revient à la diagonaliser en prenant les valeurs propres dans le sens décroissant par le théorème spectral) de la matrice $M = \frac{1}{n}\sum_{i=1}^{n} y_i^2 \mathbf{x}_i \mathbf{x}_i^T$ on obtient $\hat{\mathbf{B}}$ comme matrice de la base dans laquelle M est diagonale.

On peut alors se poser deux questions : pourquoi M permet de construire un estimateur $\hat{\mathbf{B}}$ de $\mathbf{B}$ et avec quelle précision le fait-elle ?

3.1.1 Un retour sur la preuve du Lemme 2

Le lemme 2 nous dit, qu'en supposant les hypothèses 1 et 2 décrites à 2.2 et en supposant en plus que les $\mathbf{x}_i$ sont gaussiens (loi normale d'espérance nulle et de variance égale à l'identité) on a :

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} y_i^2 \mathbf{x}_i \mathbf{x}_i^T\right] = 2\overline{\Gamma} + (1 + \mathrm{tr}(\overline{\Gamma}))\mathrm{I}_d$$

Avec $\overline{\Gamma} = \frac{1}{n}\sum_{i=1}^{n} \Gamma_i$ et $\Gamma_i = \mathbf{B}\boldsymbol{\alpha}_t \boldsymbol{\alpha}_t^T \mathbf{B}^T$. On définie en plus $\overline{\Lambda} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\alpha}_t \boldsymbol{\alpha}_t^T$
Preuve :
On va commencer par remarquer que $y_i \in \mathbb{R}$ donc $y_i = y_i^T$ et en particulier $y_i^2 = y_i y_i^T$ (1) et si on a une famille $(\mathbf{z}_i)$ des vecteurs qui suivent la même loi et sont indépendants et identiquement distribués, si $\overline{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i$ alors $\mathbb{E}[\overline{\mathbf{z}}] = \mathbb{E}[\mathbf{z}]$ (2).

Maintenant en utlisant (1) :

$$\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} y_i^2 \mathbf{x}_i \mathbf{x}_i^T] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} y_i y_i^T \mathbf{x}_i \mathbf{x}_i^T] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i^T \mathbf{B}\boldsymbol{\alpha}_t \boldsymbol{\alpha}_t^T \mathbf{B}^T \mathbf{x}_i + \epsilon_i \epsilon_i^T)\mathbf{x}_i \mathbf{x}_i^T]$$

Puis en développant :

$$\mathbb{E}[(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{B}\boldsymbol{\alpha}_t \boldsymbol{\alpha}_t^T \mathbf{B}^T \mathbf{x}_i \mathbf{x}_i \mathbf{x}_i^T) + (\frac{1}{n}\sum_{i=1}^{n} \epsilon_i \epsilon_i^T \mathbf{x}_i \mathbf{x}_i^T)] = \mathbb{E}[\mathbf{x}^T(\frac{1}{n}\sum_{i=1}^{n} \mathbf{B}\boldsymbol{\alpha}_t \boldsymbol{\alpha}_t^T \mathbf{B}^T)\mathbf{x}\mathbf{x}\mathbf{x}^T] + \mathbb{E}[\epsilon\epsilon^T \mathbf{x}\mathbf{x}^T]$$

Et comme d'une part on identifie $\overline{\Gamma}$ et de l'autre on utilise (2) car on vérifie les hypothèses 1 (en particulier que les ϵ_i et $\mathbf{x}_i$ sont indépendants) et 2, on a finalement :

$$\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} y_i^2 \mathbf{x}_i \mathbf{x}_i^T] = \mathbb{E}[\mathbf{x}^T \overline{\Gamma} \mathbf{x}\mathbf{x}\mathbf{x}^T] + \mathrm{I}_d$$

On a : $\overline{\Gamma} = \mathbf{B}\overline{\Lambda}\mathbf{B}^T$ comme $\mathbf{B}$ est une matrice orthogonale. Cependant, elle n'est pas orthogonal au sens conventionnel car elle n'est pas carré. On peut éventuellement remarquer que c'est toujours possible, même si la matrice n'est pas carré de considérer que les vecteurs colonnes sont orthogonaux (ils sont tous de même dimension) et si $r < d$, on a trivialement une famille libre de r vecteurs dans $\mathbb{R}^d$ (si $r > d$, ça ne marche évidement pas, mais ici on suppose qu'on a plus de points que de tâches sinon ça n'a pas de sens.)
On peut alors considérer $\overline{\Gamma} = \begin{bmatrix}\mathbf{B} & \mathbf{B}_\perp\end{bmatrix} \begin{bmatrix}\overline{\Lambda} & 0 \\ 0 & 0\end{bmatrix} \begin{bmatrix}\mathbf{B}^T \\ \mathbf{B}_\perp^T\end{bmatrix} = \tilde{\mathbf{B}} \begin{bmatrix}\overline{\Lambda} & 0 \\ 0 & 0\end{bmatrix} \tilde{\mathbf{B}}^T$ tel que $\tilde{\mathbf{B}}$ est bien une matrice carrée et orthogonale (qui existe par le théorème de la base incomplète car nous sommes en dimension finie) qui est simplement l'expression de $\overline{\Gamma}$ dans l'espace $M_{d,d}(\mathbb{R})$ (qui contient $M_{d,r}(\mathbb{R})$) et qui va simplement ajouter des 0 au spectre de $\overline{\Gamma}$ (et donc celui de $\overline{\Lambda}$). Or $\mathbf{B}$ est de rang au plus r et on s'intéresse seulement aux r-premières valeurs propres donc ça ne change pas les propriétés de notre décomposition. On utilisera donc $\mathbf{B}$ dans la suite à la place de $\tilde{\mathbf{B}}$ sans que cela ne change quelque chose à la validité des calculs car on a simplement besoin de vérifier l'égalité des spectres et c'est chose faite.
Maintenant, une matrice orthogonale est une matrice de changement de base orthogonale, donc $\mathrm{Sp}(\overline{\Gamma}) = \mathrm{Sp}(\overline{\Lambda})$ et donc on peut dire "quelles se comportent de la même manière" au sens des transformations dans l'espace. On va donc noter σ_i leurs valeurs propres. On va noter $Q \in \mathbb{O}_d(\mathbb{R})$ la matrice de la base orthogonale dans laquelle $\overline{\Lambda}$ est diagonale.

$$\mathbb{E}[\mathbf{x}^T \overline{\Lambda} \mathbf{x}\mathbf{x}\mathbf{x}^T] = \mathbb{E}[\mathbf{x}^T Q \mathrm{Diag}(\sigma_1, ..., \sigma_r) Q^T \mathbf{x}\mathbf{x}\mathbf{x}^T]$$

Mais, il faut garder à l'esprit que $\mathbf{x}^T\overline{\Lambda}\mathbf{x}^T \in \mathbb{R}$ et donc que $\mathbf{x}^T Q \in \mathbb{R}^d$. Par conséquent, on peut écrire $\mathbf{x}^T Q \mathrm{Diag}(\sigma_1, ..., \sigma_r) Q^T \mathbf{x} = \mathbf{x}^T Q (\mathbf{x}^T Q \mathrm{Diag}(\sigma_1, ..., \sigma_r))^T$ et si on note $Q = (\mathbf{v}_1, ..., \mathbf{v}_r)$, on a $\mathbf{x}^T Q = (\mathbf{x}^T\mathbf{v}_1, ..., \mathbf{x}^T\mathbf{v}_r)$ et donc $\mathbf{x}^T Q \mathrm{Diag}(\sigma_1, ..., \sigma_r) = (\sigma_1 \mathbf{x}^T\mathbf{v}_1, ..., \sigma_r \mathbf{x}^T\mathbf{v}_r)$.
Donc comme $\mathbf{x}^T Q (\mathbf{x}^T Q \mathrm{Diag}(\sigma_1, ..., \sigma_r))^T$ est un produit scalaire de deux vecteurs réel (car $\mathbf{x}^T\sigma_i \in \mathbb{R}$) : $\mathbf{x}^T\overline{\Lambda}\mathbf{x} = \sum_{i=1}^r \sigma_i (\mathbf{x}^T\mathbf{v}_i)^2$ et donc par linéarité de l'espérance :

$$\mathbb{E}[\mathbf{x}^T \overline{\Gamma} \mathbf{x}\mathbf{x}\mathbf{x}^T] = \sum_{i=1}^{r} \sigma_i \mathbb{E}[(\mathbf{x}^T\mathbf{v}_i)^2 \mathbf{x}\mathbf{x}^T]$$

Il suffit maitenant de trouver un moyen de calculer $\mathbb{E}[(\mathbf{x}^T\mathbf{v}_i)^2 \mathbf{x}\mathbf{x}^T]$ pour conclure. Or dans l'article, on explique que le fait que les $\mathbf{x}_i$ soient iid et gaussiens nous permet de dire que la

gausienne est isotropique (aligné selon un axe du repère) et que sa covariance s'écrit $\Sigma = \sigma^2 \mathrm{I}_d$ et comme dans notre cas, la matrice de covariance est multiplié par un réel $\mathbf{x}^T\mathbf{v}_i$ qui dépend de $\mathbf{x}$ (donc qu'on ne peut pas sortir de l'espérance) et d'un vecteur d'une base orthogonal, "il suffit" de calculer $\mathbb{E}[(\mathbf{x}^T e_1)^2 \mathbf{x}\mathbf{x}^T]$ avec e_1 un vecteur de la base canonique de $\mathbb{R}^d$ et ensuite de multiplier par $\mathbf{v}_i$ ce qui aura pour effet de faire une rotation sur la distribution et de retrouver celle qui nous intérèsse. Évidement, ce n'est pas immédiat et car je ne suis pas sûr de la méthode, on va vérifier par un calcul explicite pour un $\mathbf{v}_i$ quelconque.
On a, $\mathbf{x}^T e_1 = \mathbf{x}_1 \in \mathbb{R}$:

$$\mathbb{E}[(\mathbf{x}^T e_1)^2 \mathbf{x}\mathbf{x}^T] = \mathbb{E}[x_1^2 \mathbf{x}\mathbf{x}^T] = \mathbb{E}[x_1^2 x_i x_j]_{ij}$$

$$\mathbb{E}[x_1^2 x_i x_j]_{ij} = \begin{cases} \mathbb{E}[x_1^4] = 3 & \text{si } i = j = 1 \\ \mathbb{E}[x_1^2 x_i^2] = 1 & \text{si } i = j \neq 1 \\ \mathbb{E}[x_1^2 x_i x_j] = 0 & \text{si } i \neq j \end{cases}$$

En effet, $x \sim \mathcal{N}(0, \mathrm{I}_d)$ donc $x^2 \sim \chi^2$ à 1 degré de liberté. Donc $\mathrm{Var}(x_1^2) = \mathbb{E}[x_1^4] - (\mathbb{E}[x_1^2])^2 \Leftrightarrow \mathbb{E}[x_1^4] = 2 + 1 = 3$. Pour les deux autres moments, c'est simplement de la manipulation de vecteurs indépendants ou non. Donc :

$$\mathbb{E}[(\mathbf{x}^T e_1)^2 \mathbf{x}\mathbf{x}^T] = 2 e_1 e_1^T + \mathrm{I}_d$$

Puis, on va effectuer la même méthode pour $\mathbf{v}_i$ (donc un vecteur colonne de Q quelconque) qu'on va noter $\mathbf{v}$ dans la suite pour alléger les notations (et donc v_j désignera la j-ième composante de $\mathbf{v}_i$). En utilisant la formule du multinôme de Newton :

$$(\mathbf{x}^T \mathbf{v})^2 = (\sum_{j=1}^{d} x_j v_j)^2 = \sum_{0 \leq i,j \leq d} v_i v_j x_i x_j$$

Et note comme les v_j sont des scalaires, par línéarité de l'espérance, on a :

$$\mathbb{E}[(\mathbf{x}^T \mathbf{v})^2 \mathbf{x}\mathbf{x}^T] = \sum_{0 \leq i,j \leq d} v_i v_j \mathbb{E}[x_i x_j \mathbf{x}\mathbf{x}^T]$$

On sépare alors le cas $i = j$ et par la formule du multinôme, il suffira de prendre 2 fois la somme dans l'ordre strictement croissant des lignes par exemple. ($0 \leq i < j \leq d$).
Mais le cas $i = j$ donne directement $\sum_{i=1}^{d} v_i^2 \mathbb{E}[x_i^2 \mathbf{x}\mathbf{x}^T] = \sum_{i=1}^{d} v_i^2 (2 e_i e_i^T) + \|\mathbf{v}\|^2 \mathrm{I}_d$ (il suffit de prendre $1 = i$ dans ce qu'on a fait plus haut.) Mais par le théorème spectral, car nous sommes dans des espaces euclidiens, on peut diagonaliser en base orthonormée, et quitte à en changer, on peut la prendre normée selon la norme issue du produit scalaire donc $\|\mathbf{v}\|^2 = 1$.
Ensuite, par une disjonction de cas sur la parité des moments similaire à ce qu'on a fait plus haut :

$$\mathbb{E}[(\mathbf{x}^T \mathbf{v})^2 \mathbf{x}\mathbf{x}^T] = \sum_{i=1}^{d} v_i^2 (2 e_i e_i^T) + \mathrm{I}_d + 2 \sum_{0 \leq i < j \leq d} v_i v_j e_i e_j^T$$

$$\text{Or,} \qquad 2 \sum_{0 \leq i < j \leq d} v_i v_j e_i e_j^T = 2 \mathbf{v}\mathbf{v}^T \quad \text{donc :}$$

$$\mathbb{E}[(\mathbf{x}^T \mathbf{v})^2 \mathbf{x}\mathbf{x}^T] = 2 \mathbf{v}\mathbf{v}^T + \mathrm{I}_d$$

On trouve alors, en prenant la décomposition en vecteur propre ci-dessus dans l'autre sens (en effet, si σ_i sont des valeurs propres et $\mathbf{v}_i$ les vecteurs propres associés, une décomposition de la matrice se note $\sum_{i=1}^{r} \sigma_i \mathbf{v}_i \mathbf{v}_i^T$) et en sommant :

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} y_i^2 \mathbf{x}_i \mathbf{x}_i^T\right] = 2\overline{\Gamma} + \mathrm{tr}(\overline{\Gamma})\mathrm{I}_d + \mathrm{I}_d = 2\overline{\Gamma} + (1 + \mathrm{tr}(\overline{\Gamma}))\mathrm{I}_d$$

Ce qui conclut la preuve du Lemme 2.

3.1.2 Le lemme 2 et passage à une représentation de B

Maintenant qu'on a trouvé une expression utilisable de l'espérance théorique, on va montrer comment il va permettre de retomber sur $\mathbf{B}$. En posant $\gamma = 1 + \mathrm{Tr}(\overline{\Gamma}) = 1 + \mathrm{Tr}(\overline{\Lambda})$

$$\Omega = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} y_i^2 \mathbf{x}_i \mathbf{x}_i^T\right] = 2\overline{\Gamma} + (1 + \mathrm{tr}(\overline{\Gamma}))\mathrm{I}_d = 2\mathbf{B}\overline{\Lambda}\mathbf{B}^T + \gamma \mathrm{I}_d$$

Mais $\overline{\Lambda}$ est trivialement une matrice symétrique définie positive donc elle est aussi diagonalisable, donc $\exists Q \in \mathbb{O}_d(\mathbb{R})$ tel que $\overline{\Lambda} = Q D_{(\overline{\Lambda})} Q^T$ et comme vu plus haut, une matrice orthogonale et une matrice de passage dans une base orthogonale donc si Ω reste diagonal dans cette base alors la base à les mêmes propriétés :

$$\Omega = \mathbf{B} Q D_{(\overline{\Lambda})} \mathbf{B}^T Q^T + \gamma \mathrm{I}_d = \tilde{\mathbf{B}} D_{\overline{\Lambda}} \tilde{\mathbf{B}}^T + \gamma \mathrm{I}_d$$

Or, $\gamma \mathrm{I}_d$ commute avec toutes les matrices et elle est symétrique définie positive (hypothèse 2) donc on peut appliquer le théorème de réduction simultané des endomorphisme symétrique.

En particulier, comme $\gamma \mathrm{I}_d$ est déjà diagonale, alors γ est sa seule valeur propre et comme $2\overline{\Gamma}$ est déjà diagonalisée en base orthonormée, on sait que la base commune se trouve dans l'espace de $\tilde{\mathbf{B}}$, donc quitte à renommer ou normer les bases, on choisi $\tilde{\mathbf{B}}$ comme matrice de diagonalisation commune.
Comme $\gamma > 0$, ça ne change pas l'ordre des valeurs propre de Ω. Donc les r-premières valeurs propres de Ω sont $w_i = \sigma_i + \gamma$ et puis les $(d-r)$-autres sont exactements γ.
Par conséquent, diagonaliser Ω nous donne bien des vecteurs propres associés à l'espace des vecteurs de $\mathbf{B}$.

3.1.3 De la représentation de B à l'algorithme 1

Maintenant qu'on a montré que $\Omega = \tilde{\mathbf{B}} \tilde{D_{\overline{\Lambda}}} \tilde{\mathbf{B}}$, il reste à montrer que $M = \hat{\mathbf{B}} D_1 \hat{\mathbf{B}}^T$.
Cela revient à montrer que la méthode des moments permet effectivement de trouver un bon estimateur de $\mathbf{B}$. Pour ce faire, le théorème 7 garanti que l'erreur entre M et $\Omega = \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} y_i^2 \mathbf{x}_i \mathbf{x}_i^T]$ est suffisament petite pour faire fonctionner le théorème 3 (qui dit que dans ces conditions, $\hat{\mathbf{B}}$ est bien un bon estimateur de $\mathbf{B}$.)

3.2 Éléments de preuve du théorème 3 : une première borne sur $\hat{\mathbf{B}}$

Évidement, selon les donnés et la manière de calculer les différentes matrices, il se peut qu'on trouve une matrice $\hat{\mathbf{B}}$ très différente de $\mathbf{B}$ mais ça n'a pas d'importance. Ce qui compte, c'est

qu'elles représentent le même espace (ou en tout cas, on veut estimer la proximité entre les deux sous espaces décrits par les matrices, ie borner la différence entre les deux espaces), on cherche donc un moyen de comparer les espaces à partir d'une représentation par une matrice dans une certaine base.
Pour expliquer la stratégie de preuve du théorème 3, on va déjà expliquer le théorème de Davis-Kahan qui va nous permettre de comparer les sous espaces comme expliqué plus haut. En ce basant sur ce cours de Columbia University, on va relier les notations $\sin(\hat{\mathbf{B}}, \mathbf{B})$ et $\|\hat{\mathbf{B}}_\perp^T \mathbf{B}\|$.
On note $M = \Omega + E$ avec E bornée et on va les décomposer en somme d'actions sur des sous espaces orthogonaux :

$$\Omega = \mathbf{B} D_\Omega \mathbf{B}^T + \mathbf{B}_1 D_{\Omega_1} \mathbf{B}_1^T \quad \text{et} \quad M = \hat{\mathbf{B}} D_\Omega \hat{\mathbf{B}}^T + \hat{\mathbf{B}}_1 D_{\Omega_1} \hat{\mathbf{B}}_1^T$$

Avec $\mathbf{B} \in S$ tel que S est le sous espace des r-premières valeurs propres de Ω et $\mathbf{B}_1 \in S^\perp$ le complémentaire orthogonal (qui existe par le théorème spectral en dimension finie) ce qui nous permet d'avoir une décompositions selon les valeurs propres de Ω qui nous intérèsse. On peut alors montrer (cf. article en lien), que la proximités entre les sous espaces qui nous intérèsse revient à minimiser la norme $\|\hat{\mathbf{B}}_1^T \mathbf{B}\|$. Ce qui nous donne le théorème de Davis-Kahan :

$$\|\hat{\mathbf{B}}_1^T \mathbf{B}\| = \|\hat{\mathbf{B}}_\perp^T \mathbf{B}\| \leq \frac{\|\hat{\mathbf{B}}_\perp^T E \mathbf{B}\|}{\delta}$$

Avec un δ tel que les valeurs propres de D_Ω et D_{Ω_1} soient différentes. Il suffit alors de prendre δ tel que $\|E\| \leq \delta$ et comme $\hat{\mathbf{B}}_\perp^T$ et $\mathbf{B}$ sont des matrices orthogonales, on peut majorer l'inégalité par $\|E\|$ puis ensuite majorer l'erreur E, ce qui nous permet de conclure.