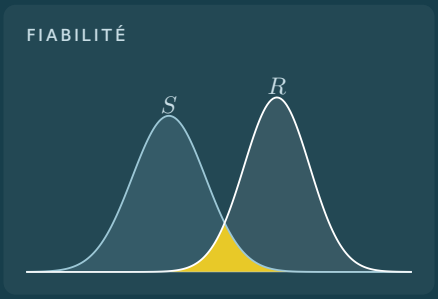
Université de Technologie de Compiègne

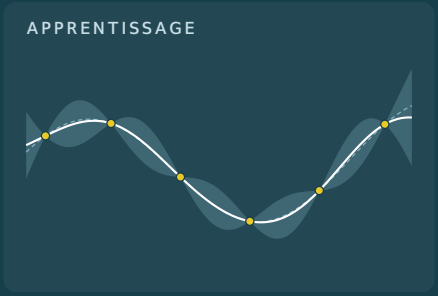
MASTER SMC

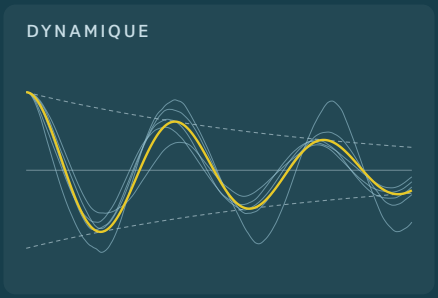
Structures et Systèmes Mécaniques Complexes (SMC)

Modélisation stochastique en mécanique

Fiabilité, Apprentissage et Dynamique







Mokhtar Z. Alaya • Salim Bouzebda

Version 2026

CHAPITRE

2 Apprentissage statistique pour la mécanique

Situation mécanique

Un essai mécanique fournit des mesures imparfaites, tandis qu'un calcul de référence peut être coûteux. On cherche à identifier les influences physiques et à prédire une réponse sans perdre le contrôle de l'erreur.

But du chapitre. Articuler régression et processus gaussiens : définir une cible mécanique, choisir un plan d'essais, évaluer une prédiction et interpréter son incertitude.

Parcours de lecture : de l'essai à une décision vérifiée. Nous conservons la poutre du chapitre précédent : mêmes appuis, mêmes positions de charges et mêmes sections critiques. Ce qui change est l'information disponible. Auparavant, l'équilibre fournissait la réponse ; ici, on imagine qu'elle doit être retrouvée à partir de mesures ou approchée à partir de quelques calculs. La réponse analytique reste accessible pour contrôler chaque étape de l'apprentissage.

La première étape consiste à préciser ce que contient une ligne de données et quelle grandeur on souhaite prédire. La deuxième cherche des coefficients dont le sens mécanique est identifiable. La troisième examine la stabilité de ces coefficients et la qualité des prédictions sur de nouvelles données. La quatrième introduit un processus gaussien pour représenter une fonction plus souple et l'incertitude qui subsiste entre les essais. Le passage d'une étape à l'autre répond donc à une difficulté concrète : définir, identifier, valider, puis quantifier ce que l'on ignore encore.

2.1 Partie A — Machine learning : apprendre une réponse mécanique

Situation mécanique

On instrumente une traverse chargée en deux points. Des capteurs mesurent les efforts et des jauges, après étalonnage, renseignent le moment dans une section. On veut retrouver les coefficients qui relient les deux charges à ce moment et vérifier qu'un capteur ne présente pas un décalage. Faut-il multiplier les mesures, changer les combinaisons de charges ou modifier le modèle ? La régression permet de répondre à ces questions si l'expérience est conçue pour distinguer les influences mécaniques.

Objectif de cette partie

Construire un jeu de données, identifier une relation linéaire, séparer apprentissage et évaluation, et comprendre la régularisation.

Quand utiliser cette partie ? Pour identifier une loi de réponse, étalonner un capteur ou construire une approximation rapide à partir d'essais ou de calculs. Le modèle de poutre fournit ici une vérité de référence : on peut séparer l'erreur de mesure, l'erreur d'identification et l'insuffisance de la famille de fonctions choisie. Cette partie ouvre la semaine 2, consacrée à l'apprentissage statistique.

La figure 2.1 donne un sens aux entrées, à la cible et aux coefficients de la régression. Les données des simulations sont synthétiques : le schéma décrit l'essai qu'elles représentent, sans prétendre à une campagne réelle.

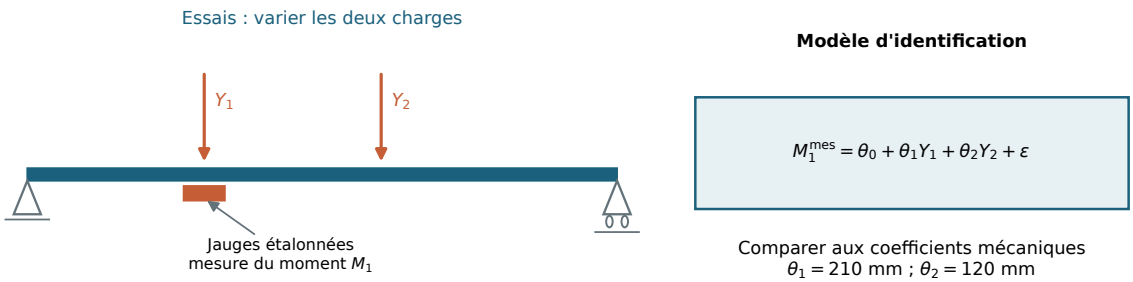


Figure 2.1 **Ce que le modèle apprend physiquement.** Des essais font varier les deux charges d'une poutre ; des jauges étalonnées à la section $x = 300$ mm permettent de déduire un moment mesuré à partir des déformations. Le modèle pédagogique est $M_1^{\text{mes}} = \theta_0 + \theta_1 Y_1 + \theta_2 Y_2 + \varepsilon$. L'équilibre donne les références $\theta_1 = 210$ mm et $\theta_2 = 120$ mm, et $\theta_0 = 0$ en l'absence de biais. Le bruit ε représente ici une erreur de mesure additive. Faire varier séparément les charges permet d'identifier leurs bras de levier effectifs ; multiplier des essais presque proportionnels apporte peu d'information pour séparer ces deux influences.

2.1.1 Quel objet apprend-on ?

Définition 2.1 (Modèle d'apprentissage supervisé).
Un jeu d'apprentissage est une collection de couples $\mathcal{D}_N = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, où $\boldsymbol{x}_i \in \mathbb{R}^p$ est une entrée et $y_i \in \mathbb{R}$ une réponse observée. Dans le modèle de régression $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$, la fonction f représente la réponse mécanique à apprendre et ε_i l'erreur d'observation. Un prédicteur $\widehat{f}$ est une fonction construite à partir de $\mathcal{D}_N$ pour estimer la réponse à une nouvelle entrée.

Dans les premiers calculs, les erreurs sont indépendantes conditionnellement aux entrées, centrées et de variance commune s_ε^2. Ce bruit de mesure ne représente pas la variabilité des charges $\boldsymbol{x}_i$. Dans la poutre, choisir $y = M_1$ donne une relation exactement linéaire. Choisir $y = \sigma_{\max}$ donne une relation affine par morceaux pour des charges positives. Le choix de la cible détermine le modèle pertinent.

2.1.1.1 Entrées, observations et domaine d'utilisation

Dans le modèle mécanique, $\boldsymbol{x}$ désigne des paramètres physiques et f la réponse du solveur. Dans l'apprentissage, les mêmes entrées deviennent les variables explicatives, et les calculs ou mesures

constituent les cibles. Apprendre un modèle signifie choisir, dans une famille de fonctions, celle dont les prédictions sont les plus adaptées au problème et aux données.

Prédire pour une nouvelle charge du même régime n'est pas la même tâche que prédire sous des charges deux fois plus grandes. Dans le premier cas, un test indépendant issu de la même loi évalue la généralisation recherchée. Dans le second, il faut tester explicitement l'extrapolation. Une séparation entraînement–test ne suffit pas si le test ne représente pas les conditions futures d'emploi.

2.1.1.2 Pourquoi minimise-t-on une erreur quadratique ?

Proposition 2.2 (Prédicteur optimal pour l'erreur quadratique).
Soit $(\boldsymbol{X}, Y)$ un couple aléatoire tel que $\mathbb{E}[Y^2] < \infty$. Notons $m(\boldsymbol{x}) = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}]$. Pour toute fonction mesurable g telle que $\mathbb{E}[g(\boldsymbol{X})^2] < \infty$, on a, pour presque toute entrée,

$$\mathbb{E}[(Y - g(\boldsymbol{x}))^2 \mid \boldsymbol{X} = \boldsymbol{x}] = \mathrm{Var}(Y \mid \boldsymbol{X} = \boldsymbol{x}) + (m(\boldsymbol{x}) - g(\boldsymbol{x}))^2. \tag{2.1}$$

La moyenne conditionnelle minimise donc le risque $\mathbb{E}[(Y - g(\boldsymbol{X}))^2]$ parmi ces prédicteurs ; le minimiseur est unique à égalité presque sûre près.

Démonstration. Écrire $Y - g(\boldsymbol{x}) = Y - m(\boldsymbol{x}) + m(\boldsymbol{x}) - g(\boldsymbol{x})$ et développer le carré. Le terme croisé s'annule car $\mathbb{E}[Y - m(\boldsymbol{x}) \mid \boldsymbol{X} = \boldsymbol{x}] = 0$. En prenant ensuite l'espérance sur $\boldsymbol{X}$, le seul terme dépendant de g est $\mathbb{E}[(m(\boldsymbol{X}) - g(\boldsymbol{X}))^2]$, minimal et nul pour $g = m$. □

Dans $Y = f(\boldsymbol{X}) + \varepsilon$, avec un bruit conditionnellement centré, cette moyenne est $f(\boldsymbol{x})$. La variance du bruit reste une erreur irréductible lorsqu'on prédit une nouvelle observation bruitée, même si f est connu exactement.

La moyenne des carrés des résidus sur l'entraînement est une version empirique de ce risque. L'optimiser adapte une fonction aux observations disponibles ; la mesurer sur de nouvelles données examine si cette adaptation se généralise.

Passage à l'identification. Une fois la cible choisie, il faut fixer la famille de réponses autorisées. Pour M_1, l'équilibre justifie une dépendance affine dans les deux charges : chercher trois coefficients a donc un fondement mécanique. Les charges sont supposées connues avec une précision suffisante ; le bruit du modèle porte sur le moment observé. Si les efforts étaient eux-mêmes fortement bruités, les moindres carrés ordinaires ne suffiraient pas à garantir des coefficients sans biais. Les équations suivantes sont ainsi des conclusions sous hypothèses, et non une procédure indépendante de la chaîne de mesure.

2.1.2 Moindres carrés et équations normales

Définition 2.3 (Matrice de régression et critère des moindres carrés).
Avec p caractéristiques et une constante, on écrit

$$\widehat{f}(\boldsymbol{x}) = \theta_0 + \sum_{j=1}^{p} \theta_j x_j, \qquad D = \begin{pmatrix} 1 & \boldsymbol{x}_1^{\mathsf{T}} \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_N^{\mathsf{T}} \end{pmatrix}. \tag{2.2}$$

La matrice $D \in \mathbb{R}^{N\times(p+1)}$ est la matrice de régression. L'estimateur des moindres carrés est un minimiseur du critère $J(\boldsymbol{\theta}) = \|\boldsymbol{y} - D\boldsymbol{\theta}\|^2$ sur $\mathbb{R}^{p+1}$.

Lemme 2.4 (Rang et matrice de Gram).
Pour toute matrice réelle D, $D^\mathsf{T}D$ est symétrique positive semi-définie et $\ker(D^\mathsf{T}D) = \ker D$. Elle est définie positive, donc inversible, si et seulement si les colonnes de D sont indépendantes.

Démonstration. Pour tout $\boldsymbol{a}$, $\boldsymbol{a}^\mathsf{T}D^\mathsf{T}D\boldsymbol{a} = \|D\boldsymbol{a}\|^2$. Cette quantité est nulle exactement lorsque $D\boldsymbol{a} = 0$, ce qui donne les affirmations sur le noyau et la positivité. □

Proposition 2.5 (Équations normales et identifiabilité).
Pour tout $D \in \mathbb{R}^{N\times(p+1)}$ et tout $\boldsymbol{y} \in \mathbb{R}^N$, le problème des moindres carrés admet au moins une solution. Ses solutions sont exactement les vecteurs vérifiant

$$D^\mathsf{T}(D\widehat{\boldsymbol{\theta}} - \boldsymbol{y}) = 0. \tag{2.3}$$

La solution est unique si et seulement si D est de rang colonne plein ; dans ce cas, $\widehat{\boldsymbol{\theta}} = (D^\mathsf{T}D)^{-1}D^\mathsf{T}\boldsymbol{y}$. Même lorsque les coefficients ne sont pas uniques, le vecteur ajusté $D\widehat{\boldsymbol{\theta}}$ est unique : c'est la projection orthogonale de $\boldsymbol{y}$ sur l'espace engendré par les colonnes de D.

L'écriture avec l'inverse est utile pour la théorie ; en calcul, utiliser une résolution par QR ou SVD, par exemple `numpy.linalg.lstsq`, plutôt qu'une inversion explicite. Des charges presque colinéaires rendent l'identification des coefficients sensible au bruit.

L'équation (2.3) signifie que le résidu est orthogonal aux colonnes de D. C'est le même principe qui conduit aux équations de Wiener–Hopf pour l'identification d'un système linéaire à partir de signaux aléatoires, repris au chapitre **??**.

2.1.2.1 Concevoir une expérience qui sépare les effets mécaniques

Si l'on charge toujours les deux points dans la même proportion, les données explorent une seule direction du plan (Y_1, Y_2). Un bon ajustement peut alors être obtenu avec plusieurs couples de coefficients : il prédit bien ce protocole, mais ne permet pas de distinguer les deux bras de levier. Pour identifier les deux influences, il faut faire varier les combinaisons de charges dans les limites physiques de l'essai.

Un capteur mal réglé peut ajouter une constante ; l'intercept permet de représenter cet effet, sans en prouver l'origine. Une structure non linéaire, un jeu dans les appuis ou un changement de condition de contact peuvent aussi produire des résidus systématiques. On confronte donc les coefficients et les résidus au modèle mécanique, au lieu de lire seulement une RMSE.

2.1.2.2 Dérivation du minimum et rôle du rang

Démonstration de la proposition 2.5. En développant le critère,

$$J(\boldsymbol{\theta}) = \boldsymbol{y}^\mathsf{T}\boldsymbol{y} - 2\boldsymbol{\theta}^\mathsf{T}D^\mathsf{T}\boldsymbol{y} + \boldsymbol{\theta}^\mathsf{T}D^\mathsf{T}D\boldsymbol{\theta}.$$

Son gradient est $2D^\mathsf{T}(D\boldsymbol{\theta} - \boldsymbol{y})$ et sa matrice hessienne est $2D^\mathsf{T}D$, positive semi-définie par le lemme 2.4. La convexité rend la condition de gradient nul nécessaire et suffisante. La projection

orthogonale de $\boldsymbol{y}$ sur $\operatorname{Im} D$ existe et possède au moins un antécédent par D : un minimiseur existe donc toujours. Si le rang est plein, la stricte convexité donne l'unicité des coefficients. Sinon, ajouter un vecteur non nul de $\ker D$ à une solution fournit une autre solution, avec le même vecteur ajusté. □

Remarque 2.6 (Identifier des coefficients ou prédire un protocole).
Des données nombreuses ne suffisent pas à garantir ce rang. Si toutes les charges expérimentales vérifient $Y_2 = 2Y_1$, répéter mille fois la même relation ne permet pas de séparer l'influence des deux charges. Il faut un plan d'expériences qui excite plusieurs directions du problème mécanique. Des prédictions précises sur ce protocole ne prouvent donc pas que chaque coefficient mécanique est identifié.

2.1.2.3 Calcul guidé : quatre essais qui séparent les deux influences

Choisissons, à titre illustratif, deux niveaux pour chaque charge de la même poutre. Les moments ci-dessous sont calculés sans bruit par l'équilibre ; ce petit tableau complète les simulations et reste distinct du jeu du TD.

Essai	Y_1 (N)	Y_2 (N)	M_1 (N mm)	z_1	z_2
1	400	200	108000	-1	-1
2	800	200	192000	1	-1
3	400	600	156000	-1	1
4	800	600	240000	1	1

Pour voir la structure du calcul, exprimons les charges par leur écart à la moyenne des essais, avec une échelle de 200 N :

$$z_1 = \frac{Y_1 - 600}{200}, \qquad z_2 = \frac{Y_2 - 400}{200}, \qquad M_1 = a_0 + a_1 z_1 + a_2 z_2.$$

Les z_j sont sans dimension ; les trois a_j sont des moments en N mm. Les colonnes constante, z_1 et z_2 sont orthogonales sur ce plan :

$$\widetilde{D} = \begin{pmatrix} 1 & -1 & -1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \qquad \widetilde{D}^{\mathsf{T}} \widetilde{D} = 4 I_3.$$

Chaque coefficient s'obtient donc par une moyenne ou un contraste entre essais, sans inversion compliquée :

$$\begin{aligned} a_0 &= \tfrac{1}{4}(108000 + 192000 + 156000 + 240000) = 174000, \\ a_1 &= \tfrac{1}{4}(-108000 + 192000 - 156000 + 240000) = 42000, \\ a_2 &= \tfrac{1}{4}(-108000 - 192000 + 156000 + 240000) = 24000. \end{aligned}$$

En redéveloppant les variables centrées, on retrouve $\theta_1 = a_1/200 = 210$ mm, $\theta_2 = a_2/200 = 120$ mm et $\theta_0 = a_0 - 3a_1 - 2a_2 = 0$ N mm. Le rang vaut trois puisque $4I_3$ est inversible. La transformation des charges n'a changé ni la poutre ni la réponse : elle a rendu le problème d'identification plus facile à lire.

Ce plan permet de modifier une charge tout en gardant l'autre constante. Le contraste qui donne a_1 compare les deux niveaux de Y_1, en moyennant sur ceux de Y_2 ; il isole ainsi la contribution de la première charge. Avec du bruit, les mêmes formules donnent des estimations qui varient d'une campagne à l'autre. La question suivante devient alors : quelle part de cette variation provient du capteur et du plan d'essais ?

Proposition 2.7 (Espérance et covariance des coefficients – complément).
Supposons le modèle affine exact $\boldsymbol{y} = D\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, D de rang colonne plein, et des erreurs de moments d'ordre deux finis telles que $\mathbb{E}[\boldsymbol{\varepsilon} \mid D] = 0$ et $\mathrm{Cov}(\boldsymbol{\varepsilon} \mid D) = s_\varepsilon^2 I$. Alors les moindres carrés sont sans biais conditionnellement au plan d'essais : $\mathbb{E}[\widehat{\boldsymbol{\theta}} \mid D] = \boldsymbol{\theta}$. De plus,

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} = (D^\mathsf{T} D)^{-1} D^\mathsf{T} \boldsymbol{\varepsilon}, \qquad \mathrm{Cov}(\widehat{\boldsymbol{\theta}} \mid D) = s_\varepsilon^2 (D^\mathsf{T} D)^{-1}.$$

La normalité des erreurs n'est pas requise pour ces identités.

Une petite valeur propre de $D^\mathsf{T} D$ produit une grande variance dans la direction correspondante.

Corollaire 2.8 (Prédire la réponse moyenne ou une nouvelle mesure).
Sous les hypothèses de la proposition 2.7, pour une entrée fixée et $\boldsymbol{d}_* = (1, \boldsymbol{x}_*^\mathsf{T})^\mathsf{T}$, on a

$$\mathrm{Var}(\boldsymbol{d}_*^\mathsf{T} \widehat{\boldsymbol{\theta}} \mid D) = s_\varepsilon^2 \boldsymbol{d}_*^\mathsf{T} (D^\mathsf{T} D)^{-1} \boldsymbol{d}_*.$$

Si $y_* = \boldsymbol{d}_*^\mathsf{T} \boldsymbol{\theta} + \varepsilon_*$ est une nouvelle mesure, avec $\mathbb{E}[\varepsilon_* \mid D] = 0$, $\mathrm{Var}(\varepsilon_* \mid D) = s_\varepsilon^2$ et une erreur indépendante des erreurs d'apprentissage conditionnellement à D, alors

$$\mathrm{Var}\left(y_* - \boldsymbol{d}_*^\mathsf{T} \widehat{\boldsymbol{\theta}} \mid D\right) = s_\varepsilon^2 \left[1 + \boldsymbol{d}_*^\mathsf{T} (D^\mathsf{T} D)^{-1} \boldsymbol{d}_*\right].$$

2.1.2.4 De la précision des coefficients à celle d'une prédiction

Démonstration de la proposition et du corollaire. Posons $B_D = (D^\mathsf{T} D)^{-1} D^\mathsf{T}$ lorsque D est de rang colonne plein. La relation $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} = B_D \boldsymbol{\varepsilon}$ explique les formules précédentes : l'espérance de l'erreur vaut zéro, et sa covariance est $B_D (s_\varepsilon^2 I) B_D^\mathsf{T}$. En utilisant $D^\mathsf{T} D$ au milieu du produit, cette expression se réduit à $s_\varepsilon^2 (D^\mathsf{T} D)^{-1}$. La variance d'une combinaison linéaire s'obtient par multiplication à gauche et à droite par $\boldsymbol{d}_*^\mathsf{T}$ et $\boldsymbol{d}_*$. Le bruit d'une nouvelle mesure étant indépendant, les deux variances de l'erreur de prédiction s'additionnent. □

La normalité du bruit n'est pas nécessaire pour ces deux résultats ; le centrage conditionnel, la covariance annoncée et le modèle affine exact le sont. Des erreurs de moyenne non nulle ou un terme mécanique oublié peuvent produire un biais malgré une résolution numérique parfaite.

Dans le plan illustratif précédent, supposons des erreurs de mesure indépendantes d'écart-type $s_\varepsilon = 1000$ N mm. Alors $\mathrm{Cov}(\widehat{\boldsymbol{a}} \mid \widetilde{D}) = (s_\varepsilon^2/4) I_3$: chaque $\widehat{a}_j$ a un écart-type de 500 N mm. Les deux pentes physiques ont chacune un écart-type de $500/200 = 2{,}5$ mm. Au centre du plan, la prédiction du moment est $\widehat{a}_0$; son écart-type est donc 500 N mm. Une nouvelle mesure indépendante y ajoute le bruit du capteur et possède une erreur de prédiction d'écart-type $\sqrt{500^2 + 1000^2} \simeq 1118$ N mm.

La constante physique mérite une lecture différente : $\widehat{\theta}_0 = \widehat{a}_0 - 3\widehat{a}_1 - 2\widehat{a}_2$. Les coefficients centrés étant non corrélés dans ce plan, $\mathrm{Var}(\widehat{\theta}_0) = 500^2(1 + 9 + 4)$, soit un écart-type d'environ 1871 N mm. Cette incertitude est plus grande qu'au centre, car la charge nulle est extérieure au rectangle exploré.

Une constante mal déterminée ne signifie donc pas nécessairement que les moments sont mal prédits au voisinage des essais : la précision dépend aussi du chargement où l'on évalue le modèle.

Pourquoi une erreur d'apprentissage n'est-elle pas toute l'erreur ?

Proposition 2.9 (Décomposition biais–variance de l'erreur).
À une entrée fixée $\boldsymbol{x}_*$, supposons que $\widehat{f}(\boldsymbol{x}_*)$ possède un moment d'ordre deux conditionnellement au plan D. Répétons les campagnes et notons $b_* = \mathbb{E}[\widehat{f}(\boldsymbol{x}_*) \mid D] - f(\boldsymbol{x}_*)$ le biais. Pour une nouvelle observation $y_* = f(\boldsymbol{x}_*) + \varepsilon_*$, dont le bruit est, conditionnellement à D, indépendant de l'apprentissage, centré et de variance s_ε^2,

$$\mathbb{E}[(y_* - \widehat{f}(\boldsymbol{x}_*))^2 \mid D] = s_\varepsilon^2 + b_*^2 + \mathrm{Var}(\widehat{f}(\boldsymbol{x}_*) \mid D). \tag{2.4}$$

Démonstration. Ajouter puis retrancher $\mathbb{E}[\widehat{f}(\boldsymbol{x}_*) \mid D]$ dans l'erreur. Après développement du carré, les termes croisés ont une espérance nulle par centrage et indépendance conditionnelle. □

Pour prédire la réponse exacte, le premier terme disparaît. Ridge peut diminuer la variance en introduisant du biais : c'est la somme pertinente pour la tâche qui doit être évaluée, ce qui motive la section suivante.

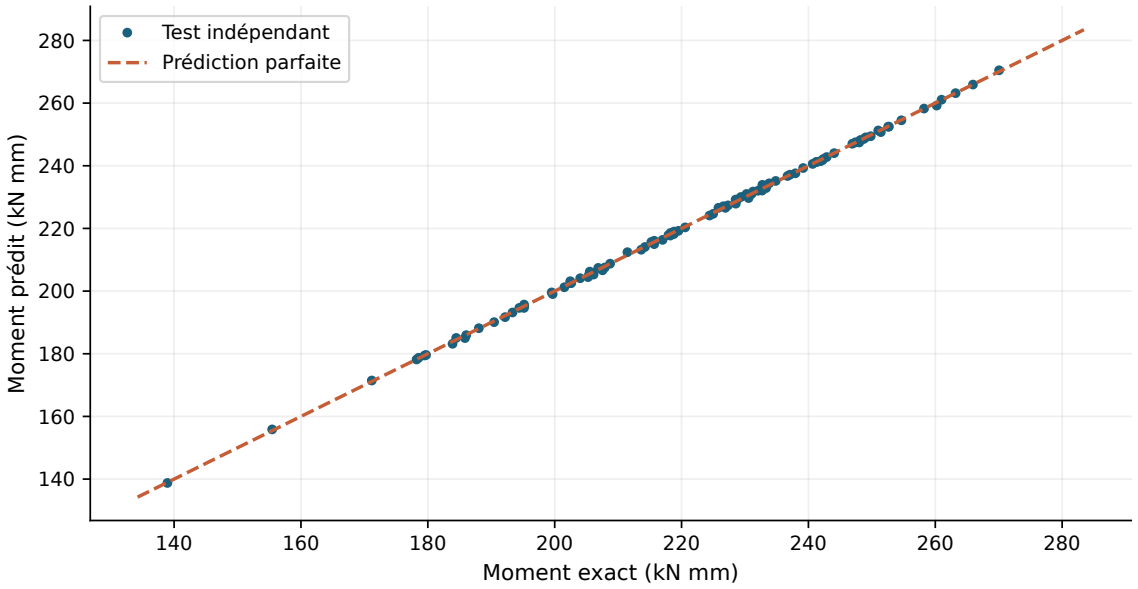


Figure 2.2 **Valider une prédiction dans l'unité mécanique.** Chaque point compare le moment exact d'un couple de charges de test au moment prédit par les moindres carrés. Ces couples sont indépendants des données d'apprentissage et la cible de test est la réponse mécanique sans bruit. La diagonale correspond à une prédiction parfaite ; un écart vertical est une erreur sur le moment, en kN mm. Un bon alignement valide cette réponse sur ce domaine de test, sans garantir des coefficients précisément identifiés ni une extrapolation correcte.

Distinguer qualité de prédiction et identification. La figure 2.3 répète l'apprentissage pour rendre visible l'incertitude des coefficients. Elle prolonge la condition de rang des équations normales par une conséquence concrète sur les essais mécaniques.

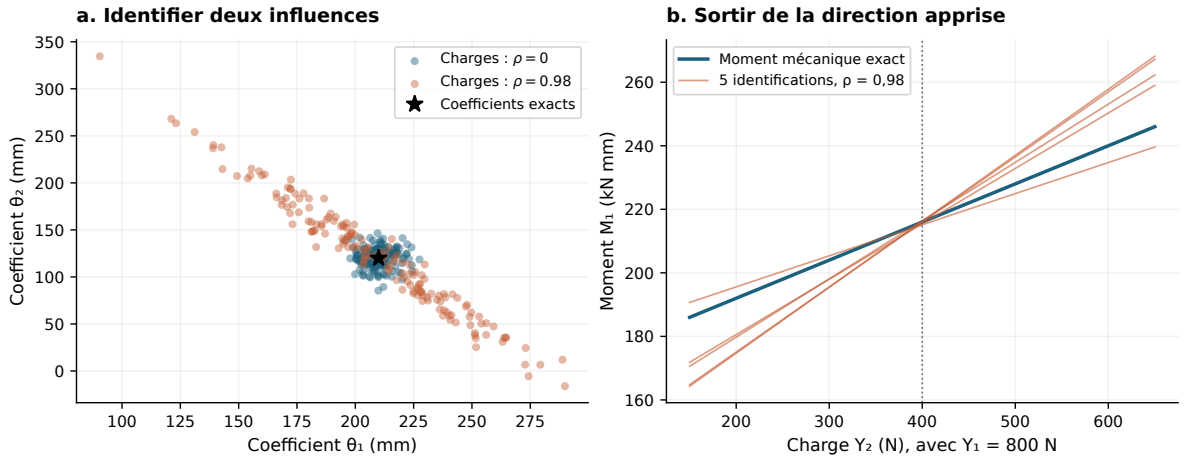


Figure 2.3 **Prédire près des essais n'est pas toujours identifier la mécanique.** Chaque point de **(a)** provient d'un ajustement sur 60 essais synthétiques ; 160 campagnes sont répétées pour chaque corrélation. La réponse est $M_1 = 210Y_1 + 120Y_2$ avec un bruit de mesure gaussien indépendant d'écart-type 5000 N mm. L'étoile indique les coefficients exacts. À $\rho = 0{,}98$, les estimations se dispersent suivant une direction allongée : un coefficient peut compenser l'autre. **(b)** Cinq ajustements de ce scénario sont comparés à $Y_1 = 800$ N. Ils s'accordent près de $Y_2 = 400$ N puis divergent quand on quitte la direction bien explorée par les charges corrélées. Ces cinq courbes ne sont pas une bande de confiance. Pour identifier deux influences physiques, le plan d'essais doit exciter deux directions indépendantes.

2.1.3 Régularisation et évaluation

Définition 2.10 (Régression ridge).
Pour un paramètre $\lambda \geq 0$, la régression ridge minimise

$$\|\boldsymbol{y} - D\boldsymbol{\theta}\|^2 + \lambda \sum_{j=1}^{p} \theta_j^2. \tag{2.5}$$

L'intercept θ_0 n'est pas pénalisé ; la matrice de pénalisation est $P = \mathrm{diag}(0, 1, \ldots, 1)$.

Proposition 2.11 (Solution ridge et unicité).
Supposons $N \geq 1$, une première colonne de D égale à $\mathbf{1}$ et $\lambda > 0$. Alors $D^\mathsf{T} D + \lambda P$ est définie positive et la solution ridge est unique, même si D n'est pas de rang colonne plein :

$$(D^\mathsf{T} D + \lambda P)\widehat{\boldsymbol{\theta}}_\lambda = D^\mathsf{T} \boldsymbol{y}.$$

Démonstration. L'annulation du gradient donne le système annoncé. Pour tout $\boldsymbol{a}$, la forme quadratique de sa matrice vaut $\|D\boldsymbol{a}\|^2 + \lambda \sum_{j=1}^{p} a_j^2$. Si elle est nulle, toutes les pentes a_j sont nulles, puis $\|a_0 \mathbf{1}\|^2 = N a_0^2 = 0$ impose $a_0 = 0$. □

Le choix de l'unité des caractéristiques modifie l'effet de la pénalisation : on centre et réduit les entrées à l'aide des statistiques de l'entraînement. Ces statistiques ne doivent pas être calculées à partir du test. Dans une validation croisée, elles sont réestimées dans chaque pli d'apprentissage.

Mettre les coefficients sur des échelles comparables. Écrivons $z_j = (Y_j - \mu_j)/s_j$ avec $s_j > 0$, où les paramètres de transformation sont fixés à partir des seuls essais d'apprentissage. Le même

modèle peut s'écrire $b_0 + \sum_j b_j z_j$, avec

$$b_j = s_j \theta_j, \qquad b_0 = \theta_0 + \sum_j \mu_j \theta_j.$$

Pour une cible en N mm, les b_j sont tous des moments. Pénaliser $\lambda \sum_{j \geq 1} b_j^2$ compare donc les effets d'une variation typique de chaque charge, plutôt que des coefficients attachés à des unités arbitraires. Avec une somme de carrés des résidus en $(\mathrm{N\,mm})^2$, λ est alors sans dimension. Dans les coordonnées physiques, la même pénalité vaut $\lambda \sum_j s_j^2 \theta_j^2$ et pondère chaque pente selon l'échelle de sa charge. Si l'on pénalise directement $\lambda_{\mathrm{phys}} \sum_j \theta_j^2$ avec des pentes en mm, λ_{phys} doit porter l'unité N^2.

Une conversion de N en kN multiplie les pentes numériques par 1000. Garder sans adaptation la même pénalité sur ces pentes changerait le problème d'optimisation, alors que la poutre n'a pas changé. La mise à l'échelle évite cet arbitraire. Elle ne rend pas indépendantes deux charges corrélées : elle traite leurs amplitudes numériques, pas le manque de directions dans le plan d'essais.

Après ajustement, on revient aux unités physiques par $\theta_j = b_j / s_j$ et $\theta_0 = b_0 - \sum_j b_j \mu_j / s_j$. L'intercept non pénalisé dans les coordonnées centrées représente la réponse au centre des essais. Si les pentes sont rétrécies, l'intercept exprimé à charge nulle peut donc changer après retour aux coordonnées physiques. Il faut toujours préciser le point auquel se rapporte une constante avant de lui attribuer une signification de précharge.

Le paramètre λ se choisit par validation interne, puis une évaluation finale est effectuée sur un jeu de test indépendant. La racine de l'erreur quadratique moyenne

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_{\mathrm{test}}} \sum_i (y_i - \widehat{f}(\boldsymbol{x}_i))^2} \tag{2.6}$$

s'exprime dans l'unité de la cible. Évaluer sur la réponse exacte $f(\boldsymbol{x})$ isole l'erreur du modèle ; évaluer sur une mesure bruitée inclut aussi le bruit de mesure.

Procédure de validation croisée

Pour des observations indépendantes, on évalue ensemble les transformations des données et le modèle de prédiction :

1. Réserver un jeu de test indépendant, puis partager les données d'apprentissage restantes en $K \geq 2$ groupes, appelés *plis*.
2. Pour chaque choix d'hyperparamètres, utiliser successivement un pli pour la validation et les $K - 1$ autres pour l'apprentissage.
3. Dans chaque découpage, estimer le centrage et la réduction sur les seuls plis d'apprentissage. Appliquer ces mêmes transformations au pli de validation, ajuster le modèle sur l'apprentissage et mesurer l'erreur de validation.
4. Agréger les erreurs des plis et choisir les hyperparamètres, par exemple λ pour ridge, sans consulter le test.
5. Réestimer les transformations et ajuster le modèle sur tout le jeu d'apprentissage avec ce choix fixé, puis effectuer l'évaluation finale sur le test réservé.

Pour des mesures groupées par pièce ou des séries temporelles, adapter le découpage à l'usage visé, comme expliqué ci-dessous.

Exemple pédagogique. Simuler 120 couples de charges corrélées, calculer M_1 puis ajouter un bruit d'écart-type 5000 N mm. Estimer les coefficients, comparer à $(210, 120)$ mm, puis évaluer sur

400 couples indépendants. Un ajustement précis des réponses n'implique pas nécessairement des coefficients précis si les variables explicatives sont très corrélées.

2.1.3.1 Exemple résolu : comprendre le retrait des coefficients

Exemple 2.12 (Une pente rétrécie par ridge).
Pour isoler le mécanisme de ridge, considérons une entrée adimensionnée $x = (-1, 0, 1)$ et des réponses $y = (-2, 1, 4)$. La réponse moyenne est $\bar{y} = 1$. Les moindres carrés retrouvent la relation $y = 1 + 3x$, puisque

$$\widehat{\theta}_1 = \frac{\sum_i x_i(y_i - \bar{y})}{\sum_i x_i^2} = \frac{6}{2} = 3.$$

Avec une pénalisation $\lambda\theta_1^2$ et $\lambda = 2$, la pente devient $6/(2+2) = 1{,}5$, tandis que l'intercept reste un. L'ajustement des trois points est moins précis. Cet exemple sans bruit montre que régulariser n'améliore pas automatiquement un modèle : la régularisation introduit un biais pour limiter la sensibilité aux données, et son utilité doit être évaluée sur la tâche visée.

Proposition 2.13 (Filtrage spectral de ridge – complément).
Après centrage des entrées et de la cible $\boldsymbol{y}_c$, soit $X = \sum_{j=1}^{r} s_j \boldsymbol{u}_j \boldsymbol{v}_j^{\mathsf{T}}$ la décomposition en valeurs singulières de la matrice des entrées, avec $s_j > 0$. Pour $\lambda > 0$, les pentes ridge sont

$$\widehat{\boldsymbol{\theta}}_{\lambda,1:p} = \sum_{j=1}^{r} \frac{s_j}{s_j^2 + \lambda} (\boldsymbol{u}_j^{\mathsf{T}} \boldsymbol{y}_c) \boldsymbol{v}_j.$$

La solution des moindres carrés de norme minimale remplace chaque facteur $s_j/(s_j^2 + \lambda)$ par $1/s_j$. Les petites valeurs singulières sont donc les plus fortement amorties relativement aux moindres carrés.

Démonstration. Dans la base des vecteurs $\boldsymbol{v}_j$, le système $(X^{\mathsf{T}}X + \lambda I)\widehat{\boldsymbol{\theta}}_{\lambda,1:p} = X^{\mathsf{T}}\boldsymbol{y}_c$ se résout composante par composante. Les composantes du noyau de X sont nulles pour $\lambda > 0$. □

2.1.3.2 Choisir le test selon l'utilisation mécanique

Pour prédire la réponse de la même traverse à de nouveaux chargements, on réserve des cas de charge indépendants. Pour prédire celle d'une nouvelle pièce, on réserve des pièces entières : répartir des mesures de la même pièce dans l'apprentissage et le test pourrait masquer la variabilité de fabrication. Pour un signal temporel, une séparation par séquences ou par campagnes évite de considérer deux instants presque identiques comme des observations nouvelles.

Le test doit aussi inclure les régions utiles au dimensionnement. Une erreur faible à petites charges renseigne peu sur le voisinage du seuil admissible. On regarde séparément les cas où la prédiction annonce « sûr » alors que le modèle de référence annonce « défaillant », car leur conséquence n'est pas la même que celle d'une alerte inutile.

2.1.3.3 Lire les erreurs avant de conclure

Une RMSE globale peut masquer une mauvaise précision près du seuil de rupture. Tracer aussi les résidus en fonction des charges et des réponses : une structure visible peut révéler une non-linéarité oubliée, une variance de bruit variable ou une région insuffisamment échantillonnée. Un modèle linéaire appliqué à $\sigma_{\max}$ peut ainsi être précis dans une région où une section domine, mais devenir insuffisant lorsque la section critique change. Pour un problème de fiabilité, les erreurs de signe de la marge au voisinage du seuil sont souvent plus pertinentes qu'une erreur moyenne sur tout le domaine.

> **À retenir**
> Un modèle appris est évalué sur des observations qui n'ont servi ni à ajuster ses coefficients, ni à sélectionner ses hyperparamètres.

Pour approfondir : James et al. [1], régression linéaire, validation et régularisation. Le lien au filtrage est développé dans les exercices de Wiener–Hopf.

De la régression à une loi sur les réponses possibles. Les moindres carrés ont séparé deux problèmes : apprendre une influence mécanique et mesurer la qualité de sa prédiction. Pour le moment M_1, la famille affine est adaptée. Pour la contrainte maximale, le changement de section critique impose un changement de pente ; augmenter le nombre d'essais ne suffit pas à rendre une droite capable de le représenter. On peut enrichir la mécanique du modèle ou choisir une famille plus souple. Le GP explore cette seconde voie et ajoute une question : entre les calculs disponibles, quelles réponses restent plausibles sous les hypothèses retenues ? Les règles de validation de la partie A restent nécessaires ; une variance fournie par le modèle ne les remplace pas.

2.2 Partie B — Machine learning : processus gaussiens

> **Situation mécanique**
> Pour dimensionner une pièce, on souhaite examiner des milliers de combinaisons de chargements. Si chaque réponse provient d'un calcul éléments finis coûteux, une simulation directe devient difficile. On apprend alors une approximation à partir d'un petit ensemble de calculs, puis on cherche où ses prédictions restent crédibles. Sur notre poutre, le passage d'une section critique à l'autre produit précisément une difficulté locale à détecter. Comment prédire la contrainte maximale sans confondre précision du modèle appris et sécurité de la structure ?

> **Objectif de cette partie**
> Construire un modèle de substitution probabiliste, interpréter son incertitude et contrôler sa qualité sur une réponse mécanique connue.

Quand utiliser cette partie ? Quand les réponses de référence sont peu nombreuses ou coûteuses et que le domaine de paramètres est maîtrisé. La poutre analytique est volontairement peu coûteuse : elle permet de contrôler les résultats avant de transférer la démarche à un solveur plus complexe. Cette partie complète la semaine 2 ; le choix du modèle mécanique reste préalable à celui du noyau statistique.

La figure 2.4 rappelle la chaîne physique que le GP doit approcher. Une faible erreur moyenne ne suffit pas si le substitut se trompe précisément près d'une contrainte admissible.

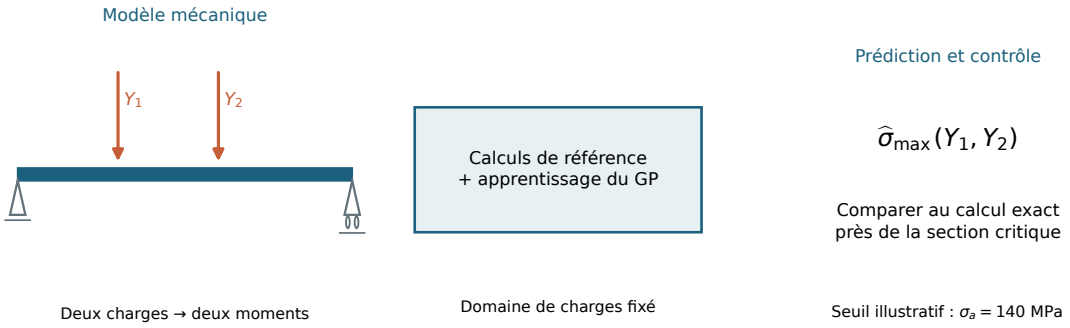


Figure 2.4 Substituer une réponse, puis contrôler son erreur. Pour chaque couple (Y_1, Y_2), l'équilibre fournit les moments des deux sections chargées, puis $\sigma_{\max} = \max(|M_1|, |M_2|)/W$. Un processus gaussien (GP) apprend cette application sur un domaine de charges annoncé. Lorsque la section qui porte le maximum change, la réponse présente un changement de pente : le modèle doit le restituer pour bien classer les cas proches d'un seuil admissible. Le seuil de 140 MPa est illustratif. La réponse analytique permet ici de vérifier le GP ; l'économie de calcul ne devient un enjeu que pour un modèle mécanique plus coûteux.

2.2.1 Une loi sur des fonctions

Définition 2.14 (Processus gaussien).
Une famille de variables aléatoires $(f(\boldsymbol{x}))_{\boldsymbol{x}\in\mathcal{X}}$ est un processus gaussien si, pour tout entier $n \geq 1$ et toute collection $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, le vecteur $(f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n))$ est conjointement gaussien, éventuellement dégénéré. On note $f \sim \mathcal{GP}(m, k)$, où

$$m(\boldsymbol{x}) = \mathbb{E}[f(\boldsymbol{x})], \qquad k(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{Cov}(f(\boldsymbol{x}), f(\boldsymbol{x}')).$$

Ces deux fonctions déterminent toutes les lois de dimension finie du processus.

On peut donc introduire le GP à partir des vecteurs gaussiens du chapitre **??** ; leur étude ne nécessite pas celle des EDS.

Définition 2.15 (Noyau de covariance admissible).
Une fonction $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est un noyau de covariance admissible si elle est symétrique et si, pour tout n, tout choix d'entrées et tout $\boldsymbol{a} \in \mathbb{R}^n$,

$$\sum_{i,j=1}^{n} a_i a_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0.$$

Ainsi toute matrice $K = (k(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j}$ est positive semi-définie. Cette propriété n'impose pas que chacun de ses coefficients soit positif, ni que K soit inversible.

Par exemple, pour $s_f^2 \geq 0$ et $\ell_j > 0$, le noyau exponentiel quadratique est

$$k(\boldsymbol{x}, \boldsymbol{x}') = s_f^2 \exp\left[-\frac{1}{2}\sum_j \frac{(x_j - x'_j)^2}{\ell_j^2}\right]. \tag{2.7}$$

Les longueurs ℓ_j contrôlent la variation selon chaque direction. Le noyau de Matérn de paramètre $3/2$, utilisé dans la démonstration enseignant, s'écrit $k = s_f^2(1+\sqrt{3}r)\exp(-\sqrt{3}r)$ avec $r^2 = \sum_j((x_j - x'_j)/\ell_j)^2$. Il impose moins de régularité que le noyau exponentiel quadratique ; cela est utile près d'un changement de section critique.

2.2.1.1 Sens du noyau et des hyperparamètres

Une grande valeur de $k(\boldsymbol{x}, \boldsymbol{x}')$ signifie que, selon l'a priori, les réponses aux deux entrées doivent fluctuer de manière proche. Le noyau décrit la similarité entre entrées du modèle, pas nécessairement une proximité dans l'espace physique. Pour la poutre, deux points du GP sont deux couples de charges, non deux sections de la poutre.

L'amplitude s_f fixe l'échelle des fluctuations de la réponse ; les longueurs ℓ_j fixent l'échelle de variation selon chaque entrée. Une longueur très grande impose une variation lente dans cette direction. Une charge exprimée en N ou en kN ne doit pas modifier la prédiction physique : les longueurs doivent changer d'unité avec les entrées. La standardisation, ajustée sur l'apprentissage uniquement, facilite cette cohérence.

Le noyau exponentiel quadratique impose des fonctions très régulières. Un noyau de Matérn permet de diminuer cette régularité. Aucun choix n'est universel : la réponse maximale de plusieurs modes mécaniques peut présenter des points anguleux, tandis qu'une réponse éloignée de toute transition peut être lisse. Une variance prédictive faible ne corrige pas un a priori de régularité inadapté.

Comprendre ce qui est aléatoire avant de conditionner.

> **Remarque 2.16** (Incertitude sur une fonction de réponse).
> Pour une géométrie et des charges imposées, le solveur de poutre fournit une réponse déterministe. Le GP est ici une représentation de notre connaissance incomplète de cette fonction. Avant les observations, plusieurs courbes sont possibles dans ce modèle ; après les observations, les courbes compatibles avec elles reçoivent davantage de poids. Répéter le calcul exact à une charge fixée ne fait donc pas varier physiquement la contrainte selon le GP.

La covariance établit comment l'information reçue à une entrée se transmet aux autres. Une observation n'agit fortement sur une prédiction que si leur covariance est significative relativement aux variances et au bruit. Le conditionnement ci-dessous rend cette idée quantitative. Avec plusieurs observations, les poids résultent aussi de leurs covariances mutuelles : ce ne sont pas nécessairement des poids positifs dont la somme vaut un. Il faut résoudre le système pour tenir compte des informations redondantes entre essais.

2.2.2 Conditionnement gaussien

> **Lemme 2.17** (Décorrélation de blocs gaussiens).
> Si les vecteurs $\boldsymbol{U}$ et $\boldsymbol{V}$ sont conjointement gaussiens, alors ils sont indépendants si et seulement si $\mathrm{Cov}(\boldsymbol{U}, \boldsymbol{V}) = 0$. La seule normalité de chaque composante séparément ne suffit pas.

Démonstration. Lorsque la covariance croisée est nulle, la matrice de covariance jointe est diagonale par blocs. La fonction caractéristique gaussienne se factorise alors en le produit des deux fonctions caractéristiques marginales, ce qui exprime l'indépendance. La réciproque résulte de la définition de la covariance. □

Théorème 2.18 (Conditionnement d'un processus gaussien).
Soit $f \sim \mathcal{GP}(0, k)$ et des observations à entrées fixées $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$, où les ε_i sont indépendants, de loi $\mathcal{N}(0, s_\varepsilon^2)$ et indépendants du processus f. Les hyperparamètres sont fixés. Posons $K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\boldsymbol{k}_* = (k(\boldsymbol{x}_i, \boldsymbol{x}_*))_i$ et $k_{**} = k(\boldsymbol{x}_*, \boldsymbol{x}_*)$. Supposons $K_y = K + s_\varepsilon^2 I$ inversible, ce qui est assuré si $s_\varepsilon^2 > 0$. Alors $f_* = f(\boldsymbol{x}_*)$ vérifie

$$\begin{pmatrix} \boldsymbol{y} \\ f_* \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} K + s_\varepsilon^2 I & \boldsymbol{k}_* \\ \boldsymbol{k}_*^\mathsf{T} & k_{**} \end{pmatrix}\right). \tag{2.8}$$

La loi conditionnelle de la *fonction latente* est $f_* \mid \boldsymbol{y} \sim \mathcal{N}(m_*, s_*^2)$, avec

$$m_* = \boldsymbol{k}_*^\mathsf{T}(K + s_\varepsilon^2 I)^{-1}\boldsymbol{y}, \tag{2.9}$$
$$s_*^2 = k_{**} - \boldsymbol{k}_*^\mathsf{T}(K + s_\varepsilon^2 I)^{-1}\boldsymbol{k}_*. \tag{2.10}$$

En particulier, $0 \leq s_*^2 \leq k_{**}$.

2.2.2.1 Démontrer le conditionnement par un résidu indépendant

Démonstration du théorème 2.18. Posons $K_y = K + s_\varepsilon^2 I$, supposée inversible, et définissons $r = f_* - \boldsymbol{k}_*^\mathsf{T} K_y^{-1}\boldsymbol{y}$. Le vecteur $(\boldsymbol{y}, r)$ est gaussien, car il est obtenu par transformation linéaire de $(\boldsymbol{y}, f_*)$. Sa covariance croisée vaut

$$\mathrm{Cov}(\boldsymbol{y}, r) = \boldsymbol{k}_* - K_y K_y^{-1}\boldsymbol{k}_* = 0.$$

Le lemme 2.17 donne l'indépendance de $\boldsymbol{y}$ et r. Ainsi, observer $\boldsymbol{y}$ ne modifie pas la loi de r, qui est centrée et de variance $k_{**} - \boldsymbol{k}_*^\mathsf{T} K_y^{-1}\boldsymbol{k}_*$. La décomposition $f_* = \boldsymbol{k}_*^\mathsf{T} K_y^{-1}\boldsymbol{y} + r$ donne les formules du cours. La variance de r est non négative, et le terme soustrait à k_{**} est non négatif puisque K_y^{-1} est définie positive. □

Avec une moyenne a priori non nulle, on applique le même calcul à $\boldsymbol{y} - \boldsymbol{m}_X$, puis on ajoute $m(\boldsymbol{x}_*)$ à la prédiction.

Corollaire 2.19 (Prédiction d'une nouvelle observation).
Sous les hypothèses du théorème 2.18, soit $y_* = f_* + \varepsilon_*$, où $\varepsilon_* \sim \mathcal{N}(0, s_{\varepsilon,*}^2)$ est indépendant de f et des erreurs d'apprentissage. Alors

$$y_* \mid \boldsymbol{y} \sim \mathcal{N}(m_*, s_*^2 + s_{\varepsilon,*}^2).$$

La variance d'une nouvelle mesure comprend donc le bruit futur, tandis que s_*^2 décrit la fonction latente conditionnellement aux observations.

Démonstration. Conditionnellement aux observations, les deux termes f_* et ε_* restent indépendants et gaussiens ; leurs moyennes et variances s'additionnent. □

Corollaire 2.20 (Interpolation sans bruit).
Si $s_\varepsilon^2 = 0$ et K est inversible, le GP retrouve exactement chaque observation : au point $\boldsymbol{x}_* = \boldsymbol{x}_i$, $m_* = y_i$ et $s_*^2 = 0$.

Démonstration. En notant $\boldsymbol{e}_i$ le i-ème vecteur de la base canonique, $\boldsymbol{k}_* = K\boldsymbol{e}_i$. Les formules du théorème donnent $m_* = \boldsymbol{e}_i^{\mathsf{T}}\boldsymbol{y} = y_i$ et $s_*^2 = K_{ii} - K_{ii} = 0$. □

Remarque 2.21 (Résolution numérique et bruit physique).
Calculer les prédictions par factorisation de Cholesky, sans inverser explicitement K_y. Un petit terme diagonal numérique, ou *jitter*, peut stabiliser la factorisation ; il ne constitue pas à lui seul un modèle de bruit physique. S'il est ajouté, les calculs portent sur la matrice perturbée et l'interpolation n'est plus exactement celle du corollaire 2.20.

2.2.2.2 Une mise à jour en MPa : rôle de la proximité et du bruit

Voici un exemple illustratif distinct des simulations du cours. À $\boldsymbol{y}_0 = (800, 600)$ N, le calcul exact de la poutre donne $\sigma_{\max} = 133{,}33$ MPa. Imaginons une mesure de 135 MPa, avec un bruit d'écart-type 5 MPa. Pour comprendre une seule mise à jour, choisissons une moyenne a priori locale de 130 MPa et une variance de 100 MPa^2. Ces paramètres sont pédagogiques, pas un étalonnage du GP.

Au point à prédire, prenons la même variance a priori et notons ρ la corrélation avec la réponse au point mesuré. La covariance croisée est alors 100ρ MPa^2, tandis que la variance de l'observation est $100 + 25 = 125$ MPa^2. Les formules donnent

$$m_* = 130 + \frac{100\rho}{125}(135 - 130), \qquad s_*^2 = 100 - \frac{(100\rho)^2}{125}.$$

Au point mesuré, $\rho = 1$: la moyenne devient 134 MPa et la variance 20 MPa^2. La mesure attire la prédiction, sans imposer exactement 135 MPa puisqu'elle est bruitée. Pour $\rho = 1/2$, on obtient 132 MPa et 80 MPa^2 ; pour $\rho \to 0$, on retrouve 130 MPa et 100 MPa^2. L'éloignement ramène ici la corrélation vers zéro pour les noyaux décroissants utilisés, et la prédiction vers son a priori.

Au point observé, lorsque la variance de bruit tend vers zéro, la moyenne tend vers la mesure et la variance latente s'annule. Lorsque le bruit devient très grand, la mesure apporte peu d'information. Pour une *nouvelle mesure* indépendante au même point, la variance vaut $20 + 25 = 45$ MPa^2, alors que celle de la réponse latente reste 20. L'exemple sans bruit qui suit permet de vérifier l'autre régime : l'interpolation exacte, cette fois avec deux observations.

2.2.2.3 Exemple résolu : interpolation de deux réponses

Exemple 2.22 (Deux observations sans bruit).
Prenons une entrée et une réponse adimensionnées, un noyau $k(x, x') = \exp(-(x - x')^2/2)$ et deux observations sans bruit $f(0) = 0$, $f(1) = 1$. Notons $c = e^{-1/2}$ et $a = e^{-1/8}$. Au milieu de l'intervalle, $x_* = 1/2$,

$$K = \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix}, \qquad \boldsymbol{k}_* = \begin{pmatrix} a \\ a \end{pmatrix}, \qquad K^{-1} = \frac{1}{1 - c^2}\begin{pmatrix} 1 & -c \\ -c & 1 \end{pmatrix}.$$

On obtient

$$m_* = \frac{a}{1 + c} \simeq 0{,}5493, \qquad s_*^2 = 1 - \frac{2a^2}{1 + c} \simeq 0{,}0305.$$

Le résultat n'est pas simplement l'interpolation affine $1/2$: il dépend du noyau et de la moyenne a priori nulle. Aux points observés, la moyenne retrouve exactement les observations et la variance est nulle, dans ce calcul sans bruit. Cet exemple permet de vérifier une implémentation avant d'utiliser un optimiseur.

2.2.2.4 Comment choisir les calculs mécaniques à apprendre ?

Le plan d'expériences doit couvrir les combinaisons de charges pertinentes. Sur la poutre, un échantillon concentré dans une zone où une seule section domine peut apprendre une branche presque linéaire et manquer le changement de section critique. Un plan plus large permet d'observer les deux régimes, mais il faut conserver un test indépendant pour vérifier l'interpolation.

Pour un problème de rupture, des calculs supplémentaires près du seuil peuvent être plus utiles que des calculs dans une zone déjà bien décrite. Cette stratégie constitue une extension d'apprentissage adaptatif ; elle ne remplace pas l'évaluation finale sur la distribution physique des charges. La distribution du plan d'apprentissage n'est pas la loi de chargement utilisée pour calculer la probabilité de défaillance.

2.2.2.5 Algorithme stable de prédiction

Former K_y, puis sa factorisation $K_y = LL^\mathsf{T}$, avec L triangulaire inférieure. Résoudre successivement

$$L\boldsymbol{z} = \boldsymbol{y}, \qquad L^\mathsf{T}\boldsymbol{\alpha} = \boldsymbol{z}, \qquad L\boldsymbol{v} = \boldsymbol{k}_*.$$

La moyenne vaut $\boldsymbol{k}_*^\mathsf{T}\boldsymbol{\alpha}$ et la variance $k_{**} - \boldsymbol{v}^\mathsf{T}\boldsymbol{v}$. Une valeur très légèrement négative, de l'ordre de l'arrondi, peut être ramenée à zéro ; une valeur nettement négative indique un problème de noyau, de calcul ou de conditionnement numérique. Le calcul dense de la factorisation coûte de l'ordre de N^3 opérations et stocke de l'ordre de N^2 coefficients. Cette approche convient donc aux petits plans d'expériences du cours.

2.2.2.6 Choix des hyperparamètres

Les hyperparamètres peuvent être choisis par maximisation de la log-vraisemblance marginale. Pour $K_y = K + s_\varepsilon^2 I$,

$$\log p(\boldsymbol{y}) = -\tfrac{1}{2}\boldsymbol{y}^\mathsf{T}K_y^{-1}\boldsymbol{y} - \tfrac{1}{2}\log\det K_y - \tfrac{N}{2}\log(2\pi). \tag{2.11}$$

Ces détails numériques sont fournis dans le notebook ; le point central est le conditionnement d'un vecteur gaussien et l'évaluation sur des points indépendants.

Proposition 2.23 (Prédiction conjointe en plusieurs points – complément).
Sous les hypothèses du théorème 2.18, soit $\boldsymbol{f}_*$ le vecteur des réponses latentes en un ensemble fini de points tests. Notons $K_{X*} = \mathrm{Cov}(\boldsymbol{y}, \boldsymbol{f}_*)$ et $K_{**} = \mathrm{Cov}(\boldsymbol{f}_*)$. La loi conditionnelle de $\boldsymbol{f}_*$ est gaussienne de moyenne $K_{X*}^\mathsf{T}K_y^{-1}\boldsymbol{y}$ et de covariance

$$\mathrm{Cov}(\boldsymbol{f}_* \mid \boldsymbol{y}) = K_{**} - K_{X*}^\mathsf{T}K_y^{-1}K_{X*}.$$

Démonstration. Le résidu vectoriel $\boldsymbol{f}_* - K_{X*}^\mathsf{T}K_y^{-1}\boldsymbol{y}$ est gaussien et non corrélé à $\boldsymbol{y}$. Appliquer le lemme 2.17, puis calculer sa covariance comme dans la démonstration scalaire. □

Sa diagonale donne les variances ponctuelles, mais les termes hors diagonale sont nécessaires pour simuler une fonction conditionnelle cohérente. Tirer indépendamment une normale à chaque point avec la bonne variance marginale ne reproduit pas les corrélations du GP.

2.2.2.7 Trois incertitudes qui répondent à trois questions différentes

La dispersion des charges répond à « quelles sollicitations peut subir la structure ? ». La variance conditionnelle du GP répond, dans son modèle, à « que sait-on de la réponse entre les calculs disponibles ? ». Le bruit d'observation répond à « avec quelle précision a-t-on mesuré ou estimé chaque réponse ? ». Ces quantités n'ont ni le même rôle ni la même unité avant propagation vers une cible commune.

Par exemple, ajouter des essais peut réduire l'incertitude du GP sans réduire la variabilité des efforts en exploitation. Inversement, renforcer la poutre modifie sa réponse physique sans garantir que le GP a été réentraîné dans ce nouveau domaine. Une bande étroite autour d'une prédiction n'est donc pas un certificat de sécurité mécanique.

2.2.2.8 Lorsque le substitut reçoit un chargement aléatoire

Pour la poutre, notons $\boldsymbol{Y} = (Y_1, Y_2)$ le futur couple de charges et $F(\boldsymbol{y})$ le GP de la contrainte maximale, exprimée en MPa. Les données d'apprentissage $\mathcal{D}$ étant fixées, posons

$$m_{\mathcal{D}}(\boldsymbol{y}) = \mathbb{E}[F(\boldsymbol{y}) \mid \mathcal{D}], \qquad s^2_{\mathcal{D}}(\boldsymbol{y}) = \mathrm{Var}(F(\boldsymbol{y}) \mid \mathcal{D}).$$

On suppose $\boldsymbol{Y}$ indépendant du processus F conditionnellement à $\mathcal{D}$, et $\mathbb{E}_{\boldsymbol{Y}|\mathcal{D}}[m_{\mathcal{D}}(\boldsymbol{Y})^2 + s^2_{\mathcal{D}}(\boldsymbol{Y})] < \infty$. La loi des charges décrit l'exploitation, et non nécessairement la répartition des points du plan d'apprentissage.

Étape 1 : fixer les charges. Grâce à l'indépendance annoncée, conditionner également par $\boldsymbol{Y} = \boldsymbol{y}$ ne modifie pas le GP. Ainsi,

$$\mathbb{E}[F(\boldsymbol{Y}) \mid \boldsymbol{Y}, \mathcal{D}] = m_{\mathcal{D}}(\boldsymbol{Y}), \qquad \mathrm{Var}(F(\boldsymbol{Y}) \mid \boldsymbol{Y}, \mathcal{D}) = s^2_{\mathcal{D}}(\boldsymbol{Y}).$$

Étape 2 : moyenner sur les chargements possibles. La loi de la variance totale, conditionnellement aux données, additionne l'espérance de la variance à charges fixées et la variance de la moyenne à charges fixées. En remplaçant ces deux quantités par les expressions précédentes, on obtient

$$\boxed{\mathrm{Var}(F(\boldsymbol{Y}) \mid \mathcal{D}) = \mathbb{E}_{\boldsymbol{Y}|\mathcal{D}}[s^2_{\mathcal{D}}(\boldsymbol{Y})] + \mathrm{Var}_{\boldsymbol{Y}|\mathcal{D}}[m_{\mathcal{D}}(\boldsymbol{Y})].} \tag{2.12}$$

Le premier terme moyenne l'incertitude sur le substitut aux chargements possibles. Le second décrit la dispersion de sa *réponse moyenne* quand les charges varient. Leur somme est une variance prédictive combinant ces deux sources ; chaque terme s'exprime en MPa^2.

Exemple chiffré et lecture mécanique. Supposons, à titre illustratif,

$$\mathbb{E}_{\boldsymbol{Y}|\mathcal{D}}[s^2_{\mathcal{D}}(\boldsymbol{Y})] = 4\ \mathrm{MPa}^2, \qquad \mathrm{Var}_{\boldsymbol{Y}|\mathcal{D}}[m_{\mathcal{D}}(\boldsymbol{Y})] = 81\ \mathrm{MPa}^2.$$

La variance totale est $4 + 81 = 85$ MPa^2, soit un écart-type $\sqrt{85} \simeq 9{,}22$ MPa. On additionne les variances, pas les écarts-types. La dispersion de la moyenne prédite domine ici, mais cela ne garantit ni la précision du GP près du seuil de rupture ni sa validité hors du domaine appris. Le mélange sur les charges n'est généralement pas gaussien : cet écart-type ne suffit pas à construire un intervalle gaussien à 95 %.

À charge fixée $\boldsymbol{y}$, seule la variance $s^2_{\mathcal{D}}(\boldsymbol{y})$ subsiste. Si la réponse mécanique exacte f est connue, cette incertitude de modèle disparaît, mais les charges aléatoires produisent encore $\mathrm{Var}_{\boldsymbol{Y}|\mathcal{D}}[f(\boldsymbol{Y})]$. Enfin, la formule concerne la contrainte latente : pour une *nouvelle observation* $Z_{\mathrm{obs}} = F(\boldsymbol{Y}) + \varepsilon$, un bruit indépendant, centré et de variance η^2 ajoute η^2 à la variance prédictive. Ce terme ne doit pas être ajouté à celle de la réponse mécanique latente.

2.2.2.9 Cibles estimées par simulation et bruit hétérogène

Si une cible est une moyenne de M_i simulations indépendantes, sa variance d'estimation peut être approchée par s_i^2/M_i, où s_i^2 est la variance empirique des réponses individuelles. Remplacer alors $s_\varepsilon^2 I$ par une matrice diagonale des variances d'estimation, lorsque les erreurs entre points d'apprentissage sont indépendantes. Des simulations utilisant les mêmes nombres aléatoires peuvent produire des erreurs corrélées : une diagonale n'est alors plus un modèle complet de leur covariance.

Si la cible est standardisée en $(y - \bar{y})/s_y$, sa variance de bruit doit être divisée par s_y^2 dans ces unités standardisées. Les variances doivent être remises à l'échelle lors de l'interprétation. Confondre variance des réponses physiques et variance de leur moyenne conduirait à surestimer le bruit des observations du GP d'un facteur voisin du nombre de simulations.

Des formules de prédiction à leur contrôle mécanique. Nous savons maintenant calculer une moyenne et une variance, et identifier la source de chaque incertitude. Il reste à vérifier ce que ces nombres disent sur la poutre. Les hyperparamètres étant traités comme fixés dans les formules, la variance conditionnelle ne quantifie pas automatiquement l'incertitude de leur estimation. Elle ne mesure pas non plus l'erreur due à un noyau inadapté au changement de section critique. C'est pourquoi la section suivante revient à une réponse analytique connue : on peut comparer séparément les valeurs prédites, leurs intervalles et la décision de dépassement. Cette vérification fournit un diagnostic concret avant de transférer la méthode à une structure dont la réponse coûte davantage à calculer.

2.2.3 Application à la poutre et limites

Le modèle apprend $(Y_1, Y_2) \mapsto \sigma_{\max}$. Pour des charges positives, $M_2 - M_1 = -90Y_1 + 120Y_2$: la section critique change sur $Y_2 = 0{,}75Y_1$. On utilise pour l'apprentissage un rectangle plus large que la zone très probable de la loi physique : $Y_1 \in [300, 1300]$, $Y_2 \in [100, 900]$ N. Le plan d'expériences et la distribution physique ont des rôles distincts.

Relier une erreur de prédiction à une erreur sur le chargement admissible. Sur la coupe $Y_1 = 800$ N, le changement de section intervient à $Y_2 = 600$ N et la contrainte vaut $133{,}33$ MPa. Au-dessus de cette charge, c'est M_2 qui commande. Avec $W = 1800$ mm^3,

$$\sigma_{\max} = \frac{96000 + 240Y_2}{1800}, \qquad Y_2 \geq 600 \text{ N}.$$

Le seuil illustratif de 140 MPa est atteint pour $Y_2 = 650$ N. Dans cette branche, une variation de 1 N de Y_2 modifie la contrainte de $240/1800 \simeq 0{,}1333$ MPa. Une erreur de réponse de 0,4 MPa correspond ainsi à environ 3 N de décalage du seuil si elle se comporte localement comme un décalage constant. Cette conversion locale donne un sens mécanique à l'erreur du substitut. Elle ne dispense pas de vérifier le changement de branche, où la pente n'est plus la même, ni de comparer les classifications pour les charges proches de 650 N.

> **Remarque 2.24** (Portée d'un intervalle crédible).
> Pour $s_* > 0$, l'intervalle $m_* \pm 1{,}96 s_*$ a une probabilité conditionnelle d'environ $95\,\%$ pour la réponse latente au point fixé. C'est un intervalle crédible *ponctuel* sous le modèle retenu et ses hyperparamètres fixés. Ce n'est ni une bande simultanée sur toutes les entrées, ni une garantie de couverture de la réponse mécanique vraie. Vérifier empiriquement les erreurs, les zones d'extrapolation et les désaccords de classification autour du seuil admissible.

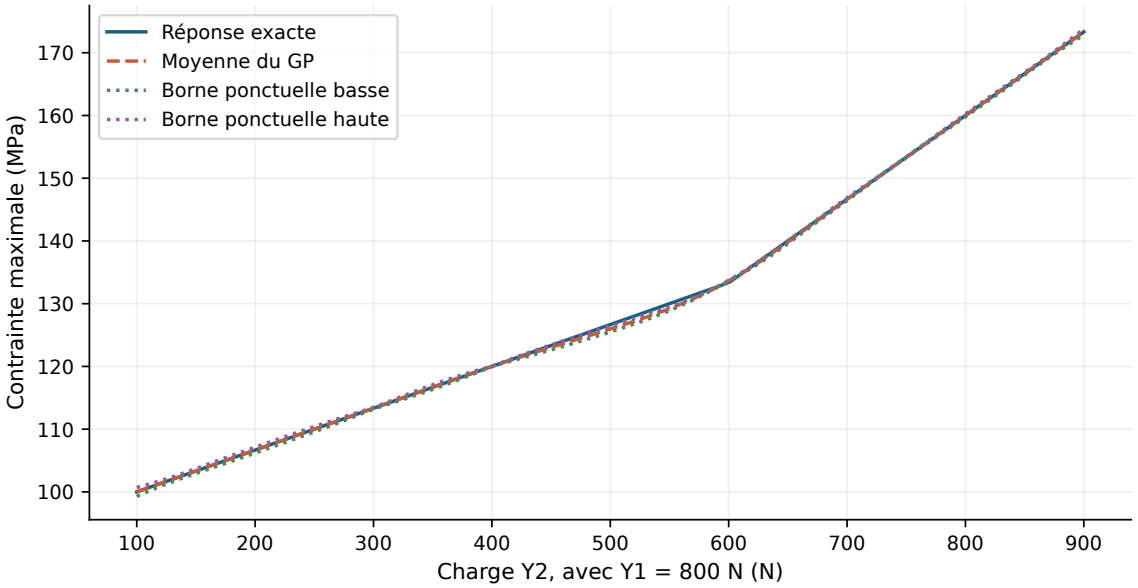


Figure 2.5 Coupe de la réponse et incertitude du GP. À $Y_1 = 800$ N, la contrainte maximale varie avec Y_2 et change de pente quand les deux sections ont le même moment. La moyenne du GP est comparée au calcul mécanique ; les bornes sont des intervalles crédibles ponctuels conditionnels au GP et à ses hyperparamètres. Leur largeur renseigne sur l'information apportée par l'apprentissage selon les charges. Elle ne représente ni la dispersion des charges physiques ni une garantie de couverture simultanée sur toute la coupe.

Comparer des budgets d'apprentissage. La figure 2.6 fixe les hyperparamètres pour isoler l'effet du nombre de calculs. Le contrôle porte sur une coupe non utilisée comme cible d'apprentissage et sur une erreur exprimée dans l'unité mécanique, le MPa.

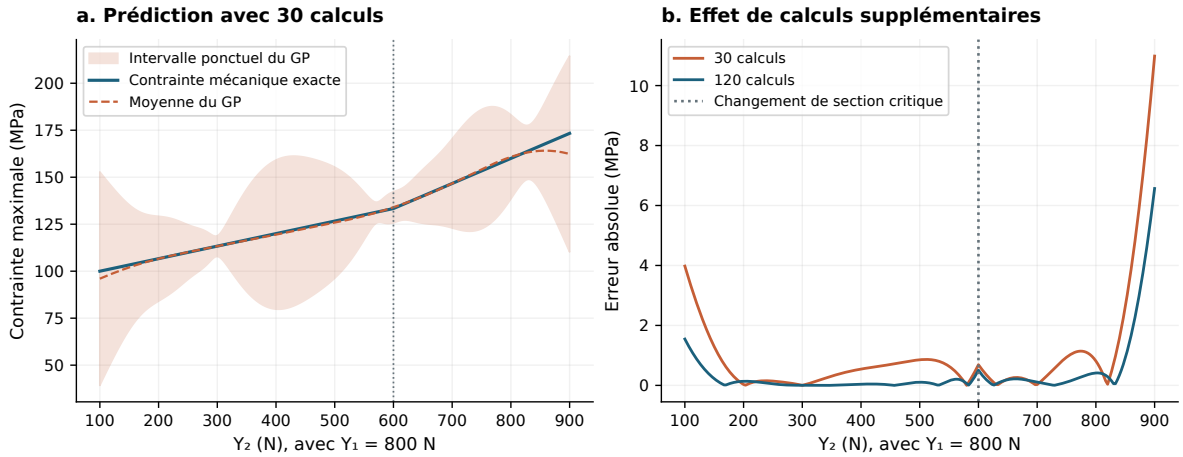


Figure 2.6 Vérifier le substitut là où la section critique change. Le GP apprend $\sigma_{\max}$ sur $[300, 1300] \times [100, 900]$ N^2 avec des ensembles emboîtés de 30 et 120 calculs exacts. Son noyau Matérn $3/2$ garde une amplitude fixe de 100 MPa et des longueurs de 0,4 dans les deux coordonnées ramenées à $[0, 1]$; aucune optimisation ni bruit d'observation n'est ajouté. **(a)** À $Y_1 = 800$ N, le changement de pente exact à $Y_2 = 600$ N correspond au changement de section critique. La bande est l'intervalle crédible ponctuel du GP à 95 %, conditionnel au modèle. **(b)** Passer de 30 à 120 calculs abaisse ici la racine de l'erreur quadratique moyenne (RMSE) sur cette coupe de $1{,}99$ à $1{,}02$ MPa, sans amélioration uniforme en chaque point. Les erreurs restent visibles près des bords : cette bande décrit l'incertitude du substitut, pas la dispersion physique des charges.

La poutre analytique est un banc de vérification ; son calcul exact est trop rapide pour revendiquer un gain de coût. L'intérêt du substitut apparaît pour des simulations plus coûteuses, comme un modèle

éléments finis ou une moyenne sur de nombreuses trajectoires. Dans ce dernier cas, la variance Monte-Carlo des cibles doit être estimée et incorporée au modèle d'observation.

Bilan mécanique avant de changer de description du chargement. La poutre relie l'identification de coefficients d'influence à la prédiction d'une contrainte maximale dont la section critique peut changer. Le plan d'essais doit distinguer les coefficients ; la validation doit porter sur les charges et les seuils utiles au dimensionnement. Les unités et le domaine d'apprentissage restent indispensables pour interpréter chaque résultat [2, chap. 2, 4 et 5].

Du chapitre 2 au chapitre 3 : que représente la fonction aléatoire ?

Le point commun est la covariance ; le changement porte sur l'objet physique et sur son indice. Le GP du présent chapitre représente une fonction de réponse encore imparfaitement connue. Le champ du chapitre suivant décrit une charge répartie qui varie d'une situation physique à l'autre.

	Réponse apprise, chapitre 2	**Charge spatiale, chapitre 3**
Fonction	$F(\boldsymbol{y})$: contrainte maximale pour des charges imposées.	$q(s)$: intensité de charge à une position de la poutre.
Indice	$\boldsymbol{y} = (y_1, y_2)$, un couple de forces en N.	$s \in [0, L]$, une abscisse en mm.
Covariance	$k_{\mathcal{D}}(\boldsymbol{y}, \boldsymbol{y}')$, entre deux situations de chargement, conditionnellement aux données.	$C_q(s, t)$, entre les intensités en deux positions lors d'une même réalisation.
Unités	$k_{\mathcal{D}}$ en MPa^2 ; longueurs du noyau en N pour des entrées non standardisées.	C_q en $(\mathrm{N/mm})^2$; longueur de corrélation en mm.
Interprétation	Incertitude du modèle sur une réponse inconnue.	Variabilité physique et dépendance spatiale du chargement.

Pourquoi reprendre la poutre ? Nous conservons les deux appuis et la portée, mais remplaçons les forces concentrées par une charge répartie $q(s)$. La fonction d'influence $a_x(s)$ mesure le moment produit à la section x par une force unitaire placée en s. La réponse devient

$$M(x) = \int_0^L a_x(s) q(s) \, \mathrm{d}s.$$

Cette écriture prolonge $\boldsymbol{M} = A\boldsymbol{Y}$: les coefficients d'une matrice deviennent les valeurs d'une fonction d'influence, et la somme devient une intégrale. Nous retrouverons exactement les coefficients 210, 120 et 240 mm dans la section **??**.

Ce qui se réutilise, ce qui doit être choisi à nouveau. Sur une grille spatiale, les valeurs du champ forment un vecteur aléatoire : les matrices de covariance et leurs factorisations restent utilisables. Si l'on choisit un *champ gaussien*, les formules de conditionnement du chapitre 2 peuvent aussi servir à estimer une charge entre des capteurs. Ce choix n'est pas obligatoire : une amplitude exponentielle ou une charge lognormale fournit un champ non gaussien. La covariance ne détermine alors pas à elle seule toutes les probabilités de dépassement.

> **À retenir**
> Un GP n'exige pas que son indice soit le temps. Dans le chapitre 2, l'indice décrit un essai ; dans le chapitre 3, il décrit notamment une position. Une petite variance du substitut ne signifie pas que la charge physique varie peu.

Bibliographie

[1] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, and Jonathan Taylor. *An Introduction to Statistical Learning : with Applications in Python*. Springer, 2023. doi : 10.1007/978-3-031-38747-0. URL https://www.statlearning.com/.

[2] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. ISBN 9780262182539. URL https://gaussianprocess.org/gpml/.