# Διαλέξεις 11 & 14: Εκτίμηση Παραμέτρων και Συμπερασματολογία κατά Bayes
# Σημειώσεις

Κωνσταντίνος Χατζηλυγερούδης
costashatz@upatras.gr

Δήμητρα Κουμούτσου
dkoumoutsou@upatras.gr

22 Δεκεμβρίου, 2025

## Contents

# 1 Κίνητρο για Εκτίμηση Παραμέτρων

Μέχρι τώρα, η εστίασή μας ήταν στην *ανάλυση σήματος*: δεδομένου ενός μετρούμενου σήματος, στόχος μας ήταν να χαρακτηρίσουμε τις ιδιότητές του μέσω φασματικής ανάλυσης, συναρτήσεων συσχέτισης και φιλτραρίσματος. Παρουσιάσαμε μη-παραμετρικά εργαλεία όπως το περιοδόγραμμα και τη μέθοδο Welch, συζητήσαμε τα φαινόμενα παραθύρωσης, τη διαρροή (leakage) και τα όρια ανάλυσης (resolution), και αναλύσαμε πώς οι στοχαστικές διεργασίες και ο θόρυβος διαδίδονται μέσω ΓΧΑ (LTI) συστημάτων καθώς και μέσω φίλτρων FIR/IIR.

Παρότι οι μέθοδοι αυτές είναι ισχυρές, οι μη-παραμετρικές περιγραφές μπορεί να καταστούν αναξιόπιστες όταν τα διαθέσιμα δεδομένα είναι μικρής διάρκειας, όταν οι μετρήσεις είναι θορυβώδεις ή χρωματισμένες, ή όταν απαιτείται υψηλή φασματική διακριτική ικανότητα. Σε τέτοιες περιπτώσεις, οι φασματικές εκτιμήσεις συχνά εμφανίζουν υψηλή διακύμανση και δεν αξιοποιούν τυχόν υποκείμενη δομή στη διαδικασία παραγωγής του σήματος.

Η παραμετρική εκτίμηση αντιμετωπίζει αυτούς τους περιορισμούς, υιοθετώντας ρητά την υπόθεση ότι το παρατηρούμενο σήμα παράγεται από ένα μοντέλο με πεπερασμένο αριθμό παραμέτρων. Αυτή η οπτική είναι ήδη έμμεσα παρούσα στο φιλτράρισμα, όπου τα σήματα μοντελοποιούνται ως έξοδος ενός συστήματος που διεγείρεται από εισόδους και θόρυβο. Η εκτίμηση των παραμέτρων τέτοιων μοντέλων μπορεί να οδηγήσει σε πιο συμπαγείς αναπαραστάσεις, αυξημένη ανθεκτικότητα στον θόρυβο, και καλύτερη γενίκευση πέρα από τα παρατηρούμενα δεδομένα.

Ο στόχος αυτού του κεφαλαίου είναι, επομένως, να μεταβούμε από την ανάλυση των σημάτων στην **εκτίμηση των μοντέλων που τα παράγουν**. Θα εισαγάγουμε οικογένειες παραμετρικών μοντέλων και θα μελετήσουμε θεμελιώδεις αρχές εκτίμησης, συμπεριλαμβανομένων των ελαχίστων τετραγώνων, της μέγιστης πιθανοφάνειας, και της Μπεϋζιανής συμπερασματολογίας.

# 2 Ελάχιστα Τετράγωνα με Γραμμικό Μοντέλο

Η εκτίμηση Ελαχίστων Τετραγώνων (Least Squares, LS) αποτελεί το βασικό εργαλείο για την προσαρμογή γραμμικών παραμετρικών μοντέλων στα δεδομένα. Υποθέτουμε ένα γραμμικό μοντέλο παρατήρησης

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{w},$$

όπου $\boldsymbol{y} \in \mathbb{R}^N$ είναι το διάνυσμα των μετρήσεων, $\boldsymbol{X} \in \mathbb{R}^{N\times p}$ είναι ένας γνωστός πίνακας σχεδίασης (ή regressor) που κατασκευάζεται από τις εισόδους και/ή από παρελθούσες εξόδους, $\boldsymbol{\theta} \in \mathbb{R}^p$ είναι το άγνωστο διάνυσμα παραμέτρων, και $\boldsymbol{w} \in \mathbb{R}^N$ αποτυπώνει τον θόρυβο μέτρησης ή/και το σφάλμα μοντελοποίησης. Η αρχή των ελαχίστων τετραγώνων επιλέγει τις παραμέτρους που ελαχιστοποιούν το συνολικό τετράγωνο του υπολοίπου (σφάλματος πρόβλεψης),

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2.$$

Γεωμετρικά, το $\boldsymbol{X}\hat{\boldsymbol{\theta}}$ είναι η ορθογώνια προβολή του $\boldsymbol{y}$ στον χώρο στηλών του $\boldsymbol{X}$, και το υπόλοιπο $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}$ είναι ορθογώνιο σε αυτόν τον υπόχωρο. Θέτοντας την παράγωγο (gradient) του κόστους ίση με μηδέν προκύπτουν οι κανονικές εξισώσεις,

$$\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\,\hat{\boldsymbol{\theta}} = \boldsymbol{X}^\mathsf{T}\boldsymbol{y}.$$

Αν ο $\boldsymbol{X}$ έχει πλήρη τάξη στηλών (οπότε ο $\boldsymbol{X}^\mathsf{T}\boldsymbol{X}$ είναι αντιστρέψιμος), ο μοναδικός ελαχιστοποιητής είναι

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}.$$

Αν ο $\boldsymbol{X}$ δεν έχει πλήρη τάξη στηλών, υπάρχουν άπειροι ελαχιστοποιητές: η τυπική επιλογή είναι η λύση ελάχιστης νόρμας, η οποία δίνεται από το Moore–Penrose ψευδοαντίστροφο,

$$\hat{\boldsymbol{\theta}} = \boldsymbol{X}^{\dagger}\boldsymbol{y}.$$

Υπό τις συνήθεις στατιστικές υποθέσεις, η LS έχει σημαντικές ιδιότητες βέλτιστου. Αν το μοντέλο είναι σωστό και $E[\boldsymbol{w}] = \boldsymbol{0}$, τότε η $\hat{\boldsymbol{\theta}}$ είναι αμερόληπτη. Αν το $\boldsymbol{w}$ είναι i.i.d. Γκαουσιανό με συνδιακύμανση $\sigma^2\boldsymbol{I}$, τότε η LS συμπίπτει με τον εκτιμητή μέγιστης πιθανοφάνειας, και μεταξύ όλων των γραμμικών αμερόληπτων εκτιμητών επιτυγχάνει ελάχιστη διακύμανση (θεώρημα Gauss–Markov: στην Γκαουσιανή περίπτωση είναι επίσης αποδοτική στην κλάση των αμερόληπτων εκτιμητών). Οι ιδέες αυτές αποτελούν τη βάση της γραμμικής παλινδρόμησης, της εκτίμησης μοντέλων AR/ARMA (μέσω διατυπώσεων γραμμικών ως προς τις παραμέτρους), και πολλών αλγορίθμων εκτίμησης που χρησιμοποιούνται στο φιλτράρισμα και στη μοντελοποίηση χώρου καταστάσεων.

## 2.1 Απόδειξη Ελαχίστων Τετραγώνων με Γραμμικό Μοντέλο

Ξεκινάμε από το γραμμικό μοντέλο παρατήρησης

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{w},$$

όπου το διάνυσμα

$$\boldsymbol{w} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}$$

αναπαριστά το υπόλοιπο (residual), αποτυπώνοντας τον θόρυβο μέτρησης και το σφάλμα μοντελοποίησης για μια δεδομένη επιλογή παραμέτρων $\boldsymbol{\theta}$.

Η αρχή των ελαχίστων τετραγώνων επιλέγει το διάνυσμα παραμέτρων που ελαχιστοποιεί την ενέργεια αυτού του υπολοίπου. Αυτό οδηγεί στη συνάρτηση κόστους

$$J(\boldsymbol{\theta}) = \|\boldsymbol{w}\|_2^2 = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}).$$

Αναπτύσσοντας τη τετραγωνική μορφή προκύπτει

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

Εφόσον πρόκειται για κυρτή τετραγωνική συνάρτηση ως προς το $\boldsymbol{\theta}$, το ελάχιστό της βρίσκεται θέτοντας την παράγωγο (gradient) ως προς $\boldsymbol{\theta}$ ίση με μηδέν:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = -2\,\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} + 2\,\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{0}.$$

Από εδώ προκύπτουν οι κανονικές εξισώσεις

$$\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}\,\hat{\boldsymbol{\theta}} = \boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}.$$

Αν ο $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ είναι αντιστρέψιμος (ισοδύναμα, αν ο $\boldsymbol{X}$ έχει πλήρη τάξη στηλών), τότε η μοναδική λύση ελαχίστων τετραγώνων είναι

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}.$$

Μια σημαντική γεωμετρική ερμηνεία προκύπτει άμεσα από τις κανονικές εξισώσεις. Ορίζοντας το εκτιμώμενο υπόλοιπο

$$\hat{\boldsymbol{w}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}},$$

παίρνουμε τη συνθήκη ορθογωνιότητας

$$\boldsymbol{X}^{\mathsf{T}}\hat{\boldsymbol{w}} = \boldsymbol{0}.$$

Έτσι, το υπόλοιπο των ελαχίστων τετραγώνων είναι ορθογώνιο στον χώρο στηλών του $\boldsymbol{X}$, πράγμα που σημαίνει ότι το προσαρμοσμένο διάνυσμα $\boldsymbol{X}\hat{\boldsymbol{\theta}}$ αποτελεί την ορθογώνια προβολή του $\boldsymbol{y}$ πάνω σε αυτόν τον υπόχωρο.

## 2.2 Εκτίμηση Ελαχίστων Τετραγώνων των Παραμέτρων Μοντέλου AR

Θεωρούμε ένα αυτοπαλίνδρομο μοντέλο τάξης $p$,

$$x[k] = -\sum_{i=1}^{p} a_i\, x[k-i] + w[k],$$

όπου $w[k]$ είναι το σφάλμα πρόβλεψης ενός βήματος μπροστά (residual). Οι άγνωστες παράμετροι είναι οι συντελεστές AR, οι οποίοι συλλέγονται στο διάνυσμα

$$\boldsymbol{a} = [a_1, \ldots, a_p]^{\mathsf{T}}.$$

Ένα βασικό σημείο είναι ότι το μοντέλο AR είναι *γραμμικό ως προς τις παραμέτρους* $a_i$. Για κάθε χρονικό δείκτη $k$ μπορούμε να γράψουμε μία γραμμική εξίσωση παλινδρόμησης

$$x[k] = -\boldsymbol{a}^{\mathsf{T}}\boldsymbol{x}_{\text{past}}[k] + w[k], \qquad \boldsymbol{x}_{\text{past}}[k] = \begin{bmatrix} x[k-1] \\ x[k-2] \\ \vdots \\ x[k-p] \end{bmatrix}.$$

Έτσι, κάθε δείγμα $x[k]$ (για $k \geq p$) παρέχει μία γραμμική εξίσωση ως προς το άγνωστο διάνυσμα $\boldsymbol{a}$.

Για να εκτιμήσουμε το $\boldsymbol{a}$ από μία σειρά δειγμάτων $\{x[0], x[1], \ldots, x[N-1]\}$, στοιβάζουμε τις εξισώσεις για $k = p, \ldots, N-1$ σε μορφή πίνακα–διανύσματος:

$$\boldsymbol{y} = -\boldsymbol{X}\boldsymbol{a} + \boldsymbol{w},$$

με

$$\boldsymbol{y} = \begin{bmatrix} x[p] \\ x[p+1] \\ \vdots \\ x[N-1] \end{bmatrix}, \qquad \boldsymbol{X} = \begin{bmatrix} x[p-1] & x[p-2] & \cdots & x[p-p] \\ x[p] & x[p-1] & \cdots & x[p-p+1] \\ \vdots & \vdots & \ddots & \vdots \\ x[N-2] & x[N-3] & \cdots & x[N-1-p] \end{bmatrix}, \qquad \boldsymbol{w} = \begin{bmatrix} w[p] \\ w[p+1] \\ \vdots \\ w[N-1] \end{bmatrix}.$$

Η εκτίμηση ελαχίστων τετραγώνων ελαχιστοποιεί το άθροισμα των τετραγώνων των σφαλμάτων πρόβλεψης ενός βήματος:

$$\hat{\boldsymbol{a}} = \arg\min_{\boldsymbol{a}} \|\boldsymbol{y} + \boldsymbol{X}\boldsymbol{a}\|_2^2.$$

Η λύση των αντίστοιχων κανονικών εξισώσεων δίνει (όταν ο $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ είναι αντιστρέψιμος)

$$\hat{\boldsymbol{a}} = -(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y},$$

και, γενικότερα, $\hat{\boldsymbol{a}} = -\boldsymbol{X}^{\dagger}\boldsymbol{y}$ χρησιμοποιώντας το ψευδοαντίστροφο.

Είναι χρήσιμο να το αντιπαραβάλουμε αυτό με την προσέγγιση Yule–Walker (YW). Η LS εκτιμά τους συντελεστές προσαρμόζοντας απευθείας τις εξισώσεις πρόβλεψης δείγμα-προς-δείγμα, ενώ η YW εκτιμά τις αυτοσυσχετίσεις και στη συνέχεια λύνει ένα Toeplitz γραμμικό σύστημα που προκύπτει από τις υποθέσεις του AR μοντέλου. Όταν τα δεδομένα ακολουθούν πράγματι ένα μοντέλο AR($p$) και είναι στάσιμα κατά ευρεία έννοια (wide-sense stationary), οι δύο προσεγγίσεις είναι ασυμπτωτικά ισοδύναμες, αλλά για μικρής διάρκειας καταγραφές ή για ήπια ασυμφωνία μοντέλου, η LS μπορεί να είναι πιο ακριβής, επειδή αποφεύγει το ενδιάμεσο βήμα της εκτίμησης των αυτοσυσχετίσεων.

# 3 Υπερπροσαρμογή και Κανονικοποίηση

Στη γραμμική παλινδρόμηση, η αύξηση του αριθμού των παραμέτρων (ή η χρήση ενός πολύ «πλούσιου» συνόλου παλινδρομητών) αυξάνει την ευελιξία του μοντέλου. Παρότι αυτό μπορεί να μειώσει το σφάλμα εκπαίδευσης, αυξάνει επίσης τον κίνδυνο της *υπερπροσαρμογής* (overfitting): η λύση των ελαχίστων τετραγώνων μπορεί να αρχίσει να προσαρμόζεται όχι μόνο στην υποκείμενη δομή του σήματος, αλλά και στον θόρυβο που υπάρχει στις μετρήσεις. Τυπικά συμπτώματα περιλαμβάνουν συντελεστές μεγάλου μέτρου, έντονη ευαισθησία σε μικρές διαταραχές των δεδομένων, και ασταθείς προβλέψεις. Ως αποτέλεσμα, ένα μοντέλο που προσαρμόζεται εξαιρετικά καλά στα διαθέσιμα δεδομένα μπορεί παρ' όλα αυτά να γενικεύει φτωχά σε νέα δεδομένα, παράγοντας μη ρεαλιστική ή έντονα ταλαντωτική συμπεριφορά.

Ένας καθιερωμένος τρόπος για τον μετριασμό της υπερπροσαρμογής είναι η εισαγωγή *κανονικοποίησης* (regularization), η οποία επιβάλλει ρητά ποινή σε υπερβολικά πολύπλοκες λύσεις. Η πιο συνηθισμένη επιλογή στα γραμμικά μοντέλα είναι η κανονικοποίηση *ridge* (ή *Tikhonov*), όπου προσθέτουμε μία ποινή $\ell_2$ στο διάνυσμα παραμέτρων:

$$\hat{\boldsymbol{\theta}}_{\text{reg}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_2^2, \qquad \lambda > 0.$$

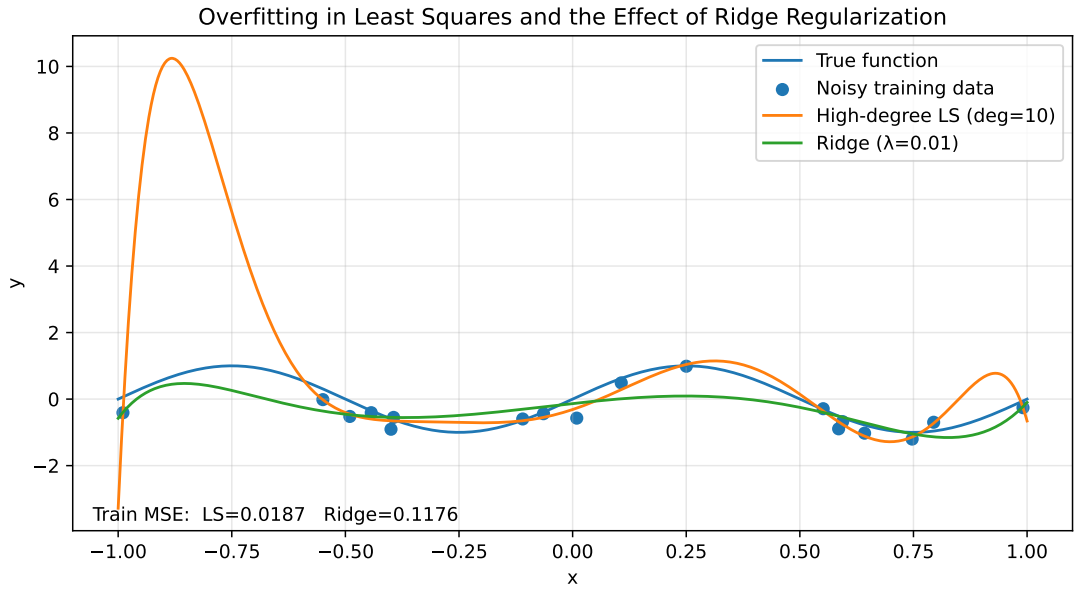
Η παράμετρος $\lambda$ ελέγχει τον συμβιβασμό μεταξύ προσαρμογής στα δεδομένα και διατήρησης μικρού μέτρου στο διάνυσμα παραμέτρων. Για $\lambda \to 0$ η λύση τείνει στα κλασικά ελάχιστα τετράγωνα, ενώ για μεγαλύτερες τιμές του $\lambda$ οι συντελεστές «συρρικνώνονται» όλο και περισσότερο προς το μηδέν.

Η ridge παλινδρόμηση επιτρέπει κλειστή μορφή λύσης:

$$\hat{\boldsymbol{\theta}}_{\text{reg}} = \left(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}.$$

Η κανονικοποίηση έχει αρκετές σημαντικές επιδράσεις. Πρώτον, μειώνει τη διακύμανση του εκτιμητή συρρικνώνοντας τα μεγέθη των παραμέτρων, κάτι που συχνά βελτιώνει την απόδοση πρόβλεψης όταν τα δεδομένα είναι θορυβώδη ή περιορισμένα. Δεύτερον, βελτιώνει τη αριθμητική σταθερότητα: ακόμη κι αν ο $\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}$ είναι κακώς υπο-συντεθειμένος (ill-conditioned) ή ιδιάζων, η προσθήκη του $\lambda\boldsymbol{I}$ καθιστά το σύστημα καλύτερα υπο-συντεθειμένο και συνήθως αντιστρέψιμο. Τέλος, η κανονικοποίηση μειώνει την ευαισθησία στον θόρυβο και σε μικρά σφάλματα μοντελοποίησης στους παλινδρομητές.

Το Σχήμα 3.1 απεικονίζει το φαινόμενο σε ένα απλοποιημένο παράδειγμα: ένα πολυώνυμο υψηλού βαθμού που προσαρμόζεται με κλασικά ελάχιστα τετράγωνα ταλαντώνεται έντονα ώστε να παρεμβάλει θορυβώδη δείγματα, ενώ η κανονικοποίηση ridge παράγει μια ομαλότερη καμπύλη που αποτυπώνει καλύτερα την υποκείμενη συνάρτηση.




Σχήμα 3.1: Υπερπροσαρμογή στα κλασικά ελάχιστα τετράγωνα (πολυώνυμο υψηλού βαθμού) και μετριασμός μέσω κανονικοποίησης ridge. Η LS προσαρμογή τείνει να «ταιριάζει» τον θόρυβο και ταλαντώνεται μεταξύ των δειγμάτων, ενώ η ridge συρρικνώνει τους συντελεστές και οδηγεί σε ένα πιο σταθερό μοντέλο με καλύτερη γενίκευση.

# 4 Μη-Γραμμικά Ελάχιστα Τετράγωνα

## 4.1 Εκτίμηση Παραμέτρων ARMA μέσω Ελαχιστοποίησης του Σφάλματος Πρόβλεψης

Τα μοντέλα αυτοπαλίνδρομου–κινούμενου μέσου (Autoregressive–Moving Average, ARMA) επεκτείνουν τα μοντέλα AR εισάγοντας ένα μέρος κινούμενου μέσου (MA), το οποίο αποτυπώνει τη δομή συσχέτισης στα υπόλοιπα (residuals). Μια διεργασία ARMA$(p, q)$ μπορεί να γραφεί ως

$$x[k] = -\sum_{i=1}^{p} a_i\, x[k-i] + \sum_{j=1}^{q} b_j\, w[k-j] + w[k], \qquad w[k] : \text{καινοτομία/λευκός θόρυβος.}$$

Το άγνωστο διάνυσμα παραμέτρων είναι

$$\boldsymbol{\theta} = \begin{bmatrix} a_1 & \cdots & a_p & b_1 & \cdots & b_q \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^{p+q}.$$

**Γιατί το πρόβλημα γίνεται μη-γραμμικό.** Για ένα καθαρό μοντέλο AR$(p)$, οι παλινδρομητές $x[k-1], \ldots, x[k-p]$ είναι γνωστοί από τα δεδομένα, άρα το μοντέλο είναι γραμμικό ως προς τους άγνωστους συντελεστές και μπορεί να γραφεί στη μορφή $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{w}$, δίνοντας λύση LS σε κλειστή μορφή. Στο ARMA, το μέρος MA εξαρτάται από παρελθούσες καινοτομίες $w[k-j]$, οι οποίες δεν είναι παρατηρήσιμες. Αν προσπαθήσουμε να ξαναγράψουμε το μοντέλο ως προς μετρήσιμες ποσότητες, προκύπτει μια αναδρομή για το σφάλμα πρόβλεψης ενός βήματος (εκτίμηση καινοτομίας). Ξεκινώντας από

$$w[k] = x[k] + \sum_{i=1}^{p} a_i\, x[k-i] - \sum_{j=1}^{q} b_j\, w[k-j],$$

ορίζουμε το σφάλμα πρόβλεψης (υπόλοιπο) ως συνάρτηση των παραμέτρων,

$$\varepsilon(k, \boldsymbol{\theta}) = x[k] + \sum_{i=1}^{p} a_i \, x[k-i] - \sum_{j=1}^{q} b_j \, \varepsilon(k-j, \boldsymbol{\theta}),$$

με κατάλληλες αρχικές συνθήκες για το $\varepsilon(k, \boldsymbol{\theta})$ για $k < 0$ (συχνά τίθεται μηδέν) και με πρώτο χρήσιμο δείκτη

$$k_0 = \max(p, q),$$

ώστε όλα τα απαιτούμενα παρελθόντα δείγματα και παρελθόντα σφάλματα να είναι διαθέσιμα. Το κρίσιμο σημείο είναι ότι το $\varepsilon(k, \boldsymbol{\theta})$ εξαρτάται από προηγούμενα $\varepsilon(k-j, \boldsymbol{\theta})$, τα οποία με τη σειρά τους εξαρτώνται από το $\boldsymbol{\theta}$. Επομένως, η ακολουθία υπολοίπων είναι *μη-γραμμική συνάρτηση* των παραμέτρων και δεν μπορούμε να διατυπώσουμε το πρόβλημα ως ένα ενιαίο γραμμικό σύστημα $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}$.

**Στόχος μη-γραμμικών ελαχίστων τετραγώνων.** Μία τυπική προσέγγιση είναι η *ελαχιστοποίηση του σφάλματος πρόβλεψης* (prediction error minimization), η οποία επιλέγει παραμέτρους που ελαχιστοποιούν το άθροισμα των τετραγώνων των σφαλμάτων πρόβλεψης ενός βήματος:

$$J(\boldsymbol{\theta}) = \frac{1}{2} \sum_{k=k_0}^{N-1} \varepsilon(k, \boldsymbol{\theta})^2.$$

Ορίζοντας το στοιβαγμένο διάνυσμα υπολοίπων

$$\boldsymbol{\varepsilon}(\boldsymbol{\theta}) = \begin{bmatrix} \varepsilon(k_0, \boldsymbol{\theta}) \\ \varepsilon(k_0+1, \boldsymbol{\theta}) \\ \vdots \\ \varepsilon(N-1, \boldsymbol{\theta}) \end{bmatrix} \in \mathbb{R}^M, \qquad M = N - k_0,$$

μπορούμε να γράψουμε

$$J(\boldsymbol{\theta}) = \frac{1}{2} \, \boldsymbol{\varepsilon}(\boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{\varepsilon}(\boldsymbol{\theta}) = \frac{1}{2} \, \big\| \boldsymbol{\varepsilon}(\boldsymbol{\theta}) \big\|_2^2, \qquad \hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}).$$

Σε αντίθεση με τη γραμμική LS, αυτό το πρόβλημα βελτιστοποίησης γενικά δεν έχει λύση σε κλειστή μορφή και πρέπει να λυθεί επαναληπτικά.

## 4.2 Gauss–Newton για Μη-Γραμμικά Ελάχιστα Τετράγωνα

Η μέθοδος Gauss–Newton είναι μια κλασική μέθοδος ελαχιστοποίησης συναρτήσεων ελαχίστων τετραγώνων. Προχωρά τοπικά γραμμικοποιώντας το διάνυσμα υπολοίπων γύρω από την τρέχουσα εκτίμηση $\boldsymbol{\theta}_i$. Έστω ότι ο Ιακωβιανός του διανύσματος υπολοίπων είναι

$$\boldsymbol{J}_{\varepsilon}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\varepsilon}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{M \times P}, \qquad P = p + q, \qquad \big[\boldsymbol{J}_{\varepsilon}(\boldsymbol{\theta})\big]_{m,\ell} = \frac{\partial \varepsilon(k_0 + m, \boldsymbol{\theta})}{\partial \theta_\ell}.$$

Μια προσέγγιση Taylor πρώτης τάξης δίνει

$$\boldsymbol{\varepsilon}(\boldsymbol{\theta}_i + \Delta\boldsymbol{\theta}) \approx \boldsymbol{\varepsilon}(\boldsymbol{\theta}_i) + \boldsymbol{J}_{\varepsilon}(\boldsymbol{\theta}_i) \, \Delta\boldsymbol{\theta}.$$

Η Gauss–Newton επιλέγει τότε $\Delta\boldsymbol{\theta}_i$ ως τον ελαχιστοποιητή του προκύπτοντος γραμμικού υποπροβλήματος ελαχίστων τετραγώνων:

$$\Delta\boldsymbol{\theta}_i = \arg\min_{\Delta\boldsymbol{\theta}} \|\boldsymbol{\varepsilon}(\boldsymbol{\theta}_i) + \boldsymbol{J}_\varepsilon(\boldsymbol{\theta}_i)\,\Delta\boldsymbol{\theta}\|_2^2 .$$

Αυτό οδηγεί στις κανονικές εξισώσεις

$$\boldsymbol{J}_\varepsilon(\boldsymbol{\theta}_i)^{\mathsf{T}}\boldsymbol{J}_\varepsilon(\boldsymbol{\theta}_i)\,\Delta\boldsymbol{\theta}_i = -\boldsymbol{J}_\varepsilon(\boldsymbol{\theta}_i)^{\mathsf{T}}\boldsymbol{\varepsilon}(\boldsymbol{\theta}_i),$$

και ακολουθεί η ενημέρωση των παραμέτρων

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \Delta\boldsymbol{\theta}_i,$$

η οποία μπορεί προαιρετικά να συνδυαστεί με απόσβεση (damping) ή αναζήτηση γραμμής (line search) για βελτίωση της ευρωστίας.

**Ερμηνεία.** Σε κάθε επανάληψη, η Gauss–Newton αντικαθιστά την μη-γραμμική απεικόνιση των υπολοίπων με την τοπική γραμμική της προσέγγιση και λύνει ένα γραμμικό πρόβλημα ελαχίστων τετραγώνων. Η μέθοδος έχει επίσης μια χρήσιμη ερμηνεία δεύτερης τάξης: για στόχους ελαχίστων τετραγώνων,

$$\nabla J(\boldsymbol{\theta}) = \boldsymbol{J}_\varepsilon(\boldsymbol{\theta})^{\mathsf{T}}\boldsymbol{\varepsilon}(\boldsymbol{\theta}), \qquad \nabla^2 J(\boldsymbol{\theta}) = \boldsymbol{J}_\varepsilon^{\mathsf{T}}\boldsymbol{J}_\varepsilon + \sum_{m=1}^{M} \varepsilon_m(\boldsymbol{\theta})\,\nabla^2\varepsilon_m(\boldsymbol{\theta}),$$

και η Gauss–Newton προσεγγίζει τον Εσσιανό παραλείποντας τον δεύτερο όρο, δίνοντας

$$\nabla^2 J(\boldsymbol{\theta}) \approx \boldsymbol{J}_\varepsilon(\boldsymbol{\theta})^{\mathsf{T}}\boldsymbol{J}_\varepsilon(\boldsymbol{\theta}),$$

κάτι που αποφεύγει τον υπολογισμό δεύτερων παραγώγων ενώ αξιοποιεί τη δομή των ελαχίστων τετραγώνων.

## 4.3 Γενική Μορφή Gauss–Newton

Γενικότερα, έστω υπόλοιπα $\boldsymbol{r}(\boldsymbol{\theta}) \in \mathbb{R}^M$ και αντικειμενική συνάρτηση

$$J(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{r}(\boldsymbol{\theta})\|_2^2 = \boldsymbol{r}(\boldsymbol{\theta})^T\boldsymbol{r}(\boldsymbol{\theta}),$$

με Ιακωβιανό

$$\boldsymbol{J}_r(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{r}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{M\times P},$$

η Gauss–Newton υπολογίζει το $\Delta\boldsymbol{\theta}_i$ από

$$\boldsymbol{J}_r(\boldsymbol{\theta}_i)^{\mathsf{T}}\boldsymbol{J}_r(\boldsymbol{\theta}_i)\,\Delta\boldsymbol{\theta}_i = -\boldsymbol{J}_r(\boldsymbol{\theta}_i)^{\mathsf{T}}\boldsymbol{r}(\boldsymbol{\theta}_i), \qquad \boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \Delta\boldsymbol{\theta}_i.$$

Στην εκτίμηση ARMA, το $\boldsymbol{r}(\boldsymbol{\theta})$ αντιστοιχεί στα στοιβαγμένα σφάλματα πρόβλεψης $\boldsymbol{\varepsilon}(\boldsymbol{\theta})$, και η μέθοδος βελτιώνει επαναληπτικά τους συντελεστές ώστε να ελαχιστοποιήσει την ενέργεια του σφάλματος πρόβλεψης ενός βήματος.

## 4.4 Παράδειγμα Gauss–Newton: Εκτίμηση Μοντέλου ARMA(1,1)

Θεωρούμε το μοντέλο ARMA(1,1)

$$x[k] = -a\,x[k-1] + b\,w[k-1] + w[k], \qquad w[k] \text{ καινοτομία.}$$

Επειδή οι παρελθούσες καινοτομίες $w[k-1]$ δεν είναι παρατηρήσιμες, εργαζόμαστε με το *σφάλμα πρόβλεψης ενός βήματος* (εκτίμηση καινοτομίας)

$$\varepsilon(k, \boldsymbol{\theta}) = x[k] + a\,x[k-1] - b\,\varepsilon(k-1, \boldsymbol{\theta}), \qquad \boldsymbol{\theta} = \begin{bmatrix} a \\ b \end{bmatrix},$$

με μια τυπική αρχικοποίηση όπως $\varepsilon(0, \boldsymbol{\theta}) = 0$. Το κόστος μη-γραμμικών ελαχίστων τετραγώνων (σφάλμα πρόβλεψης) είναι

$$J(\boldsymbol{\theta}) = \frac{1}{2}\sum_{k=1}^{N-1} \varepsilon(k, \boldsymbol{\theta})^2 = \frac{1}{2}\big\|\boldsymbol{\varepsilon}(\boldsymbol{\theta})\big\|_2^2, \qquad \boldsymbol{\varepsilon}(\boldsymbol{\theta}) = \begin{bmatrix} \varepsilon(1, \boldsymbol{\theta}) \\ \vdots \\ \varepsilon(N-1, \boldsymbol{\theta}) \end{bmatrix}.$$

Εφόσον το $\varepsilon(k, \boldsymbol{\theta})$ εξαρτάται αναδρομικά από παρελθόντα σφάλματα, είναι μη-γραμμική συνάρτηση του $(a, b)$, και ελαχιστοποιούμε το $J(\boldsymbol{\theta})$ επαναληπτικά χρησιμοποιώντας Gauss–Newton.

**Ιακωβιανός μέσω αναδρομών.** Η Gauss–Newton απαιτεί τον Ιακωβιανό του διανύσματος υπολοίπων:

$$\boldsymbol{J}_{\varepsilon}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\varepsilon}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \partial_a \varepsilon(1, \boldsymbol{\theta}) & \partial_b \varepsilon(1, \boldsymbol{\theta}) \\ \vdots & \vdots \\ \partial_a \varepsilon(N-1, \boldsymbol{\theta}) & \partial_b \varepsilon(N-1, \boldsymbol{\theta}) \end{bmatrix}.$$

Παραγωγίζοντας την αναδρομή $\varepsilon(k) = x[k] + a\,x[k-1] - b\,\varepsilon(k-1)$ προκύπτουν εμπρόσθιες αναδρομές για τις μερικές παραγώγους:

$$\partial_a \varepsilon(k) = x[k-1] - b\,\partial_a \varepsilon(k-1),$$
$$\partial_b \varepsilon(k) = -\varepsilon(k-1) - b\,\partial_b \varepsilon(k-1),$$

με αρχικές συνθήκες

$$\varepsilon(0) = 0, \qquad \partial_a \varepsilon(0) = 0, \qquad \partial_b \varepsilon(0) = 0.$$

Έτσι, για μια δεδομένη αρχική εικασία παραμέτρων $(a, b)$, μπορούμε να υπολογίσουμε $\varepsilon(k)$, $\partial_a \varepsilon(k)$, και $\partial_b \varepsilon(k)$ σε ένα μόνο εμπρόσθιο πέρασμα μέσα από τα δεδομένα.

**Επανάληψη Gauss–Newton.** Σε κάθε επανάληψη $i$, δεδομένου $\boldsymbol{\theta}_i = [a_i \;\; b_i]^{\mathsf{T}}$, υπολογίζουμε $\boldsymbol{\varepsilon}(\boldsymbol{\theta}_i)$ και $\boldsymbol{J}_{\varepsilon}(\boldsymbol{\theta}_i)$, και στη συνέχεια λύνουμε το γραμμικό βήμα ελαχίστων τετραγώνων

$$\Delta\boldsymbol{\theta}_i = \arg\min_{\Delta\boldsymbol{\theta}} \big\|\boldsymbol{\varepsilon}(\boldsymbol{\theta}_i) + \boldsymbol{J}_{\varepsilon}(\boldsymbol{\theta}_i)\Delta\boldsymbol{\theta}\big\|_2^2,$$

που δίνει τις κανονικές εξισώσεις

$$\boldsymbol{J}_{\varepsilon}(\boldsymbol{\theta}_i)^{\mathsf{T}} \boldsymbol{J}_{\varepsilon}(\boldsymbol{\theta}_i)\,\Delta\boldsymbol{\theta}_i = -\boldsymbol{J}_{\varepsilon}(\boldsymbol{\theta}_i)^{\mathsf{T}} \boldsymbol{\varepsilon}(\boldsymbol{\theta}_i).$$

Η ενημέρωση είναι

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \Delta\boldsymbol{\theta}_i,$$

και σταματάμε όταν το $\|\Delta\boldsymbol{\theta}_i\|$ (ή η μείωση του $J$) είναι μικρό. Στην πράξη, μπορεί να προστεθεί απόσβεση/αναζήτηση κατά μήκος αν το απλό βήμα είναι ασταθές.

## 4.5 Παράδειγμα (Μία επανάληψη Gauss–Newton)

Παίρνουμε τρία δείγματα:

$$x[0] = 1, \quad x[1] = 0.5, \quad x[2] = -0.2,$$

και αρχικοποιούμε με

$$\boldsymbol{\theta}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Rightarrow \quad a_0 = 0,\ b_0 = 0, \qquad \varepsilon(0) = 0.$$

**Βήμα 1: υπόλοιπα στο $\boldsymbol{\theta}_0$.** Χρησιμοποιώντας $\varepsilon(k) = x[k] + a\,x[k-1] - b\,\varepsilon(k-1)$ με $a_0 = b_0 = 0$:

$$\varepsilon(1, \boldsymbol{\theta}_0) = x[1] = 0.5, \qquad \varepsilon(2, \boldsymbol{\theta}_0) = x[2] = -0.2.$$

Άρα

$$\boldsymbol{\varepsilon}(\boldsymbol{\theta}_0) = \begin{bmatrix} 0.5 \\ -0.2 \end{bmatrix}.$$

**Βήμα 2: Ιακωβιανός στο $\boldsymbol{\theta}_0$.** Με $b_0 = 0$, οι αναδρομές των παραγώγων απλοποιούνται σε

$$\partial_a \varepsilon(k) = x[k-1], \qquad \partial_b \varepsilon(k) = -\varepsilon(k-1).$$

Επομένως,

$$\partial_a \varepsilon(1) = x[0] = 1, \qquad \partial_a \varepsilon(2) = x[1] = 0.5,$$
$$\partial_b \varepsilon(1) = -\varepsilon(0) = 0, \qquad \partial_b \varepsilon(2) = -\varepsilon(1) = -0.5.$$

Άρα,

$$\boldsymbol{J}_{\varepsilon}(\boldsymbol{\theta}_0) = \begin{bmatrix} 1 & 0 \\ 0.5 & -0.5 \end{bmatrix}.$$

**Βήμα 3: Βήμα Gauss–Newton.** Υπολογίζουμε

$$\boldsymbol{J}^{\mathsf{T}}\boldsymbol{J} = \begin{bmatrix} 1.25 & -0.25 \\ -0.25 & 0.25 \end{bmatrix}, \qquad \boldsymbol{J}^{\mathsf{T}}\boldsymbol{\varepsilon} = \begin{bmatrix} 0.4 \\ 0.1 \end{bmatrix}.$$

Εφόσον $\det\big(\boldsymbol{J}^{\mathsf{T}}\boldsymbol{J}\big) = 0.25$, ο αντίστροφος είναι

$$(\boldsymbol{J}^{\mathsf{T}}\boldsymbol{J})^{-1} = \begin{bmatrix} 1 & 1 \\ 1 & 5 \end{bmatrix}.$$

Άρα,

$$\Delta\boldsymbol{\theta}_0 = -(\boldsymbol{J}^{\mathsf{T}}\boldsymbol{J})^{-1}\boldsymbol{J}^{\mathsf{T}}\boldsymbol{\varepsilon} = -\begin{bmatrix} 1 & 1 \\ 1 & 5 \end{bmatrix}\begin{bmatrix} 0.4 \\ 0.1 \end{bmatrix} = \begin{bmatrix} -0.5 \\ -0.9 \end{bmatrix},$$

και οι ενημερωμένες παράμετροι είναι

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta}_0 = \begin{bmatrix} -0.5 \\ -0.9 \end{bmatrix}.$$

Αυτό δείχνει τη μηχανική της Gauss–Newton: υπολογίζουμε υπόλοιπα και Ιακωβιανό μέσω εμπρόσθιων αναδρομών, και στη συνέχεια λύνουμε ένα μικρό γραμμικό σύστημα για να ενημερώσουμε τις παραμέτρους.

## 4.6 Πρακτική Οπτική: Μη-Γραμμικά Ελάχιστα Τετράγωνα μέσα από το Πρίσμα της Βελτιστοποίησης

Τα μη-γραμμικά ελάχιστα τετράγωνα αποτελούν ένα πρόβλημα βελτιστοποίησης της μορφής

$$\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}), \qquad J(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{r}(\boldsymbol{\theta})\|_2^2,$$

όπου $\boldsymbol{r}(\boldsymbol{\theta}) \in \mathbb{R}^M$ είναι το διάνυσμα υπολοίπων (π.χ. τα στοιβαγμένα σφάλματα πρόβλεψης ARMA). Από τη σκοπιά της γενικής βελτιστοποίησης, η Gauss–Newton απλώς κατασκευάζει, σε κάθε επανάληψη $\boldsymbol{\theta}_i$, ένα τοπικό τετραγωνικό μοντέλο του $J$ βασισμένο σε μια γραμμικοποίηση πρώτης τάξης του $\boldsymbol{r}$ και στη συνέχεια πραγματοποιεί ένα βήμα καθόδου.

**Παράγωγος και (προσεγγιστικός) Εσσιανός.** Χρησιμοποιώντας τον κανόνα αλυσίδας,

$$\nabla J(\boldsymbol{\theta}) = \boldsymbol{J}_r(\boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{r}(\boldsymbol{\theta}), \qquad \boldsymbol{J}_r(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{r}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

Ο ακριβής Εσσιανός είναι

$$\nabla^2 J(\boldsymbol{\theta}) = \boldsymbol{J}_r(\boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{J}_r(\boldsymbol{\theta}) + \sum_{m=1}^{M} r_m(\boldsymbol{\theta})\, \nabla^2 r_m(\boldsymbol{\theta}),$$

και η Gauss–Newton χρησιμοποιεί την προσέγγιση

$$\boldsymbol{H}_{\mathrm{GN}}(\boldsymbol{\theta}) \;\triangleq\; \boldsymbol{J}_r(\boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{J}_r(\boldsymbol{\theta}),$$

η οποία είναι ακριβής όταν τα υπόλοιπα είναι μικρά κοντά στο βέλτιστο (ή όταν το μοντέλο είναι κοντά στο σωστό).

**Υπολογισμός βήματος ως λύση γραμμικού συστήματος.** Η διεύθυνση Gauss–Newton $\Delta\boldsymbol{\theta}_i$ ικανοποιεί

$$\boldsymbol{H}_{\mathrm{GN}}(\boldsymbol{\theta}_i)\, \Delta\boldsymbol{\theta}_i = -\nabla J(\boldsymbol{\theta}_i) = -\boldsymbol{J}_r(\boldsymbol{\theta}_i)^{\mathsf{T}} \boldsymbol{r}(\boldsymbol{\theta}_i).$$

Αυτό είναι άμεσα ανάλογο με ένα βήμα Newton, αλλά με έναν προσεγγιστικό Εσσιανό που αποφεύγει τις δεύτερες παραγώγους και αξιοποιεί τη δομή των ελαχίστων τετραγώνων.

**Αναζήτηση γραμμής (globalization).** Όπως και στη γενική μη-γραμμική βελτιστοποίηση, το πλήρες βήμα $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \Delta\boldsymbol{\theta}_i$ μπορεί να αποτύχει όταν η αρχική υποθέση ή όταν το τετραγωνικό μοντέλο δεν είναι ακριβές. Μία τυπική αντιμετώπιση είναι η *αναζήτηση γραμμής* (line search):

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \alpha_i\, \Delta\boldsymbol{\theta}_i, \qquad \alpha_i \in (0, 1],$$

όπου το $\alpha_i$ επιλέγεται ώστε να εξασφαλίζει επαρκή μείωση του $J$ (π.χ. συνθήκη Armijo) και να διατηρεί τις επαναλήψεις σταθερές. Στην πράξη, μια αναζήτηση γραμμής με οπισθοχώρηση (backtracking) είναι συχνά επαρκής: ξεκινάμε από $\alpha_i = 1$ και μειώνουμε επαναληπτικά $\alpha_i \leftarrow \rho\alpha_i$ (με $\rho \in (0,1)$) έως ότου το $J(\boldsymbol{\theta}_{i+1})$ μειώνεται επαρκώς.

**Απόσβεση (Levenberg–Marquardt).** Μια άλλη ευρέως χρησιμοποιούμενη σταθεροποίηση είναι η *απόσβεση* (damping), που οδηγεί στο βήμα Levenberg–Marquardt (LM):

$$\big(\boldsymbol{J}_r(\boldsymbol{\theta}_i)^\mathsf{T}\boldsymbol{J}_r(\boldsymbol{\theta}_i) + \mu_i \boldsymbol{I}\big)\,\Delta\boldsymbol{\theta}_i = -\boldsymbol{J}_r(\boldsymbol{\theta}_i)^\mathsf{T}\boldsymbol{r}(\boldsymbol{\theta}_i), \qquad \mu_i > 0.$$

Για μεγάλο $\mu_i$, το βήμα συμπεριφέρεται σαν κατάβαση κλίσης (μικρό, συντηρητικό, ευρωστό)[1]; για μικρό $\mu_i$, προσεγγίζει τη Gauss–Newton (γρήγορη τοπική σύγκλιση). Αυτό συνδέεται στενά με την οπτική των περιοχών εμπιστοσύνης (trust-region): ο αλγόριθμος περιορίζει τα βήματα όταν το τοπικό μοντέλο είναι αναξιόπιστο και τα μεγαλώνει όταν η προβλεπόμενη μείωση συμφωνεί με την πραγματική.

**Πρακτικός οδηγός.** Δοθέντος του τρέχοντος εκτιμητή $\boldsymbol{\theta}_i$:

1. Υπολόγισε τα υπόλοιπα $\boldsymbol{r}(\boldsymbol{\theta}_i)$ (π.χ. σφάλματα πρόβλεψης ARMA μέσω εμπρόσθιας αναδρομής).
2. Υπολόγισε τον Ιακωβιανό $\boldsymbol{J}_r(\boldsymbol{\theta}_i)$ (αναλυτικά μέσω αναδρομών, ή αριθμητικά αν χρειάζεται).
3. Υπολόγισε μια διεύθυνση αναζήτησης λύνοντας είτε:
$$\boldsymbol{J}^\mathsf{T}\boldsymbol{J}\,\Delta\boldsymbol{\theta} = -\boldsymbol{J}^\mathsf{T}\boldsymbol{r} \quad \text{(Gauss–Newton)}, \qquad \text{ή} \qquad (\boldsymbol{J}^\mathsf{T}\boldsymbol{J} + \mu\boldsymbol{I})\,\Delta\boldsymbol{\theta} = -\boldsymbol{J}^\mathsf{T}\boldsymbol{r} \quad \text{(LM)}.$$
4. Ενημέρωσε χρησιμοποιώντας μήκος βήματος $\alpha_i$ (αναζήτηση κατά μήκος) και/ή παράμετρο απόσβεσης $\mu_i$:
$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \alpha_i \Delta\boldsymbol{\theta}_i.$$
5. Σταμάτα όταν το $J(\boldsymbol{\theta}_i)$, το $\|\nabla J(\boldsymbol{\theta}_i)\|$, ή το $\|\Delta\boldsymbol{\theta}_i\|$ είναι επαρκώς μικρό.

**Τελικές παρατηρήσεις.** Από την οπτική της βελτιστοποίησης, η Gauss–Newton είναι απλώς "η μέθοδος Newton εξειδικευμένη στα ελάχιστα τετράγωνα", με μια προσέγγιση του Εσσιανού που είναι φθηνή υπολογιστικά και συχνά πολύ αποτελεσματική. Όταν συνδυάζεται με τυπικά εργαλεία globalization (αναζήτηση γραμμής ή trust-region/απόσβεση), γίνεται ένα ευρωστό και πρακτικό εργαλείο για προβλήματα εκτίμησης παραμέτρων όπως η προσαρμογή ARMA μέσω σφάλματος πρόβλεψης, η μη-γραμμική παλινδρόμηση, και πολλές εργασίες εκτίμησης παραμέτρων σε μοντέλα χώρου καταστάσεων.

# 5 Μη-Γραμμικά Αυτοπαλίνδρομα Μοντέλα και Γραμμικά Μοντέλα σε Χαρακτηριστικά

## 5.1 Μοντέλα NAR

Πολλά πραγματικά σήματα δεν μπορούν να περιγραφούν με ακρίβεια μέσω γραμμικής αυτοπαλίνδρομης δυναμικής. Φαινόμενα όπως ο κορεσμός, οι νεκρές ζώνες, η υστέρηση, οι μεταγωγές (switching) και ο θόρυβος που εξαρτάται από την κατάσταση εισάγουν μη-γραμμικές εξαρτήσεις από τις προηγούμενες τιμές. Μια φυσική επέκταση ενός μοντέλου AR($p$) είναι το *μη-γραμμικό αυτοπαλίνδρομο* (Nonlinear Autoregressive, NAR) μοντέλο

$$x[k] = f\big(x[k-1],\, x[k-2],\, \ldots,\, x[k-p]; \boldsymbol{\theta}\big) + w[k],$$

[1] Δεδομένης της προτίμησης ορολογίας, ο όρος "κατάβαση κλίσης" αντιστοιχεί στο gradient descent.

όπου $f(\cdot;\boldsymbol{\theta})$ είναι μια μη-γραμμική απεικόνιση που παραμετροποιείται από ένα άγνωστο διάνυσμα $\boldsymbol{\theta}$ και $w[k]$ συγκεντρώνει τον θόρυβο και το σφάλμα μοντελοποίησης. Η τάξη $p$ καθορίζει το μήκος μνήμης, δηλαδή πόσα παρελθόντα δείγματα επηρεάζουν την πρόβλεψη.

Από πλευράς εκτίμησης, η εκτίμηση παραμέτρων των μοντέλων NAR διατυπώνεται φυσικά ως ένα (γενικά) πρόβλημα *μη-γραμμικών ελαχίστων τετραγώνων*. Για κάθε $k \geq p$ ορίζουμε το υπόλοιπο πρόβλεψης ενός βήματος

$$\varepsilon(k,\boldsymbol{\theta}) = x[k] - f\big(x[k-1],\ldots,x[k-p];\boldsymbol{\theta}\big),$$

και εκτιμούμε τις παραμέτρους ελαχιστοποιώντας το άθροισμα των τετραγώνων των υπολοίπων:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2}\sum_{k=p}^{N-1} \varepsilon(k,\boldsymbol{\theta})^2.$$

Σε αντίθεση με τα γραμμικά AR, τυπικά δεν υπάρχει λύση σε κλειστή μορφή, οπότε χρησιμοποιούμε επαναληπτικές μεθόδους βελτιστοποίησης (κατάβαση κλίσης, Gauss–Newton, Levenberg–Marquardt). Το πρακτικό υπολογιστικό κόστος αυτών των μεθόδων κυριαρχείται από την επαναλαμβανόμενη (i) αξιολόγηση του $f(\cdot;\boldsymbol{\theta})$ σε όλο το σύνολο δεδομένων και (ii) τον υπολογισμό παραγώγων των υπολοίπων ως προς $\boldsymbol{\theta}$.

Εννοιολογικά, το NAR γενικεύει το AR αντικαθιστώντας τον γραμμικό προβλέπτη $-\sum_{i=1}^{p} a_i x[k-i]$ με μια μη-γραμμική συνάρτηση του ίδιου διανύσματος παρελθόντων δειγμάτων.

## 5.2 Γραμμικά Μοντέλα σε Χαρακτηριστικά

Ένας χρήσιμος συμβιβασμός μεταξύ γραμμικής και μη-γραμμικής μοντελοποίησης είναι να διατηρήσουμε το μοντέλο *γραμμικό ως προς τις παραμέτρους* ενώ επιτρέπουμε μη-γραμμική εξάρτηση από τις εισόδους μέσω μιας απεικόνισης χαρακτηριστικών. Έστω $\boldsymbol{x} \in \mathbb{R}^d$ ένα διάνυσμα εισόδου (ή παλινδρομητών), και ορίζουμε ένα διάνυσμα χαρακτηριστικών

$$\boldsymbol{\phi}(\boldsymbol{x}) = \begin{bmatrix} \phi_1(\boldsymbol{x}) \\ \phi_2(\boldsymbol{x}) \\ \vdots \\ \phi_P(\boldsymbol{x}) \end{bmatrix} \in \mathbb{R}^P.$$

Ένα μοντέλο γραμμικό-σε-χαρακτηριστικά είναι

$$y = \boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{\phi}(\boldsymbol{x}) + w, \qquad \boldsymbol{\theta} \in \mathbb{R}^P.$$

Δοθέντων δεδομένων $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^{N}$, στοιβάζουμε τα διανύσματα χαρακτηριστικών στον πίνακα σχεδίασης

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\phi}(\boldsymbol{x}_1)^{\mathsf{T}} \\ \boldsymbol{\phi}(\boldsymbol{x}_2)^{\mathsf{T}} \\ \vdots \\ \boldsymbol{\phi}(\boldsymbol{x}_N)^{\mathsf{T}} \end{bmatrix} \in \mathbb{R}^{N\times P}, \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}.$$

Τότε η εκτίμηση παραμέτρων ανάγεται σε κλασικά ελάχιστα τετράγωνα:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\|_2^2, \qquad \hat{\boldsymbol{\theta}} = (\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{y},$$

υποθέτοντας ότι ο $\boldsymbol{\Phi}$ έχει πλήρη τάξη στηλών (αλλιώς χρησιμοποιούμε $\boldsymbol{\Phi}^{\dagger}$ ή κανονικοποίηση ridge).

Αυτή η οπτική είναι σημαντική διότι επιτρέπει μη-γραμμική προσέγγιση συναρτήσεων, διατηρώντας ταυτόχρονα την απλότητα των εργαλείων γραμμικής εκτίμησης (LS, ridge/Tikhonov και τις στατιστικές ερμηνείες τους). Η μη-γραμμικότητα «μεταφέρεται» στην επιλογή των συναρτήσεων βάσης $\phi_i(\cdot)$, ενώ οι άγνωστες παράμετροι εξακολουθούν να εισάγονται γραμμικά.

Συνηθισμένες επιλογές χαρακτηριστικών περιλαμβάνουν πολυωνυμικές επεκτάσεις βάσης (π.χ. $[1, x, x^2, \ldots]$), τριγωνομετρικά/χαρακτηριστικά Fourier (π.χ. $\sin(\omega x)$ και $\cos(\omega x)$), ακτινικές συναρτήσεις βάσης (radial basis functions), και μαθημένα χαρακτηριστικά (π.χ. την έξοδο ενός σταθερού επιπέδου νευρωνικού δικτύου). Η πολυπλοκότητα του μοντέλου ελέγχεται από τον αριθμό και τον τύπο των χαρακτηριστικών και, στην πράξη, από την κανονικοποίηση όταν το $P$ είναι μεγάλο σε σχέση με το $N$.

## 5.3 Το NAR ως Γραμμική Παλινδρόμηση σε Χαρακτηριστικά (Ειδική Περίπτωση)

Μια διδακτική ειδική περίπτωση είναι όταν η συνάρτηση NAR εκφράζεται ως γραμμικός συνδυασμός επιλεγμένων μη-γραμμικών χαρακτηριστικών του διανύσματος παρελθόντων δειγμάτων. Ορίζουμε τον παλινδρομητή

$$\boldsymbol{z}[k] = \begin{bmatrix} x[k-1] \\ x[k-2] \\ \vdots \\ x[k-p] \end{bmatrix},$$

επιλέγουμε μια απεικόνιση χαρακτηριστικών $\boldsymbol{\phi}(\boldsymbol{z}[k])$, και μοντελοποιούμε

$$x[k] = \boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{\phi}(\boldsymbol{z}[k]) + w[k].$$

Αυτό είναι ένα μη-γραμμικό αυτοπαλίνδρομο μοντέλο ως προς τη συμπεριφορά εισόδου–εξόδου, αλλά παραμένει *γραμμικό ως προς τις άγνωστες παραμέτρους* $\boldsymbol{\theta}$, οπότε μπορούμε να το εκτιμήσουμε μέσω (κανονικοποιημένων) ελαχίστων τετραγώνων. Αυτή η οπτική συχνά αποκαλείται *NAR με συναρτήσεις βάσης* (basis-function NAR) και παρέχει μια πρακτική γέφυρα μεταξύ της εκτίμησης γραμμικών AR και της πλήρως μη-γραμμικής ταυτοποίησης NAR.

## 5.4 Λυμένα Παραδείγματα: Γραμμικά Μοντέλα σε Χαρακτηριστικά

**Παράδειγμα 1: Αφινικό μοντέλο.** Θεωρούμε ξανά την απλή απεικόνιση χαρακτηριστικών

$$\boldsymbol{\phi}(x) = \begin{bmatrix} 1 \\ x \end{bmatrix}, \qquad y = \boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{\phi}(x) + w, \qquad \boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}.$$

Υποθέτουμε ότι παρατηρούμε τα εξής τρία δείγματα:

$$(x_1, y_1) = (0, 0), \quad (x_2, y_2) = (1, 1), \quad (x_3, y_3) = (2, 3).$$

Ο πίνακας σχεδίασης και το διάνυσμα εξόδου είναι

$$\boldsymbol{\Phi} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}.$$

Υπολογίζουμε

$$\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi} = \begin{bmatrix} 3 & 3 \\ 3 & 5 \end{bmatrix}, \qquad \boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{y} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}.$$

Άρα,

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{y} = \frac{1}{6}\begin{bmatrix} 5 & -3 \\ -3 & 3 \end{bmatrix}\begin{bmatrix} 4 \\ 7 \end{bmatrix} = \begin{bmatrix} -1/6 \\ 7/6 \end{bmatrix}.$$

Το προσαρμοσμένο μοντέλο είναι επομένως

$$\hat{y}(x) = -\frac{1}{6} + \frac{7}{6}x.$$

**Παράδειγμα 2: Τετραγωνικά χαρακτηριστικά.** Τώρα αυξάνουμε την ευελιξία του μοντέλου χρησιμοποιώντας μια τετραγωνική απεικόνιση ως χαρακτηριστικό:

$$\boldsymbol{\phi}(x) = \begin{bmatrix} 1 \\ x \\ x^2 \end{bmatrix}, \qquad y = \boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{\phi}(x) + w, \qquad \boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}.$$

Θεωρούμε τέσσερα δείγματα:

$$(x_1, y_1) = (-1, 1), \quad (x_2, y_2) = (0, 0), \quad (x_3, y_3) = (1, 1), \quad (x_4, y_4) = (2, 4).$$

Ο πίνακας σχεδίασης και το διάνυσμα εξόδου είναι

$$\boldsymbol{\Phi} = \begin{bmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 4 \end{bmatrix}.$$

Υπολογίζουμε

$$\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi} = \begin{bmatrix} 4 & 2 & 6 \\ 2 & 6 & 8 \\ 6 & 8 & 18 \end{bmatrix}, \qquad \boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{y} = \begin{bmatrix} 6 \\ 8 \\ 18 \end{bmatrix}.$$

Λύνοντας $(\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi})\hat{\boldsymbol{\theta}} = \boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{y}$ παίρνουμε

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Άρα, το προσαρμοσμένο μοντέλο είναι

$$\hat{y}(x) = x^2.$$

**Συμπέρασμα.** Αλλάζοντας την απεικόνιση χαρακτηριστικών $\boldsymbol{\phi}(x)$, μπορούμε να μεταβούμε από απλά γραμμικά μοντέλα σε πιο εκφραστικούς μη-γραμμικούς προβλέπτες, ενώ εξακολουθούμε να βασιζόμαστε ακριβώς στην ίδια μηχανική των ελαχίστων τετραγώνων. Η πολυπλοκότητα του μοντέλου καθορίζεται πλήρως από τα επιλεγμένα χαρακτηριστικά (και, στην πράξη, από την κανονικοποίηση όταν αυξάνεται η διάσταση των χαρακτηριστικών).

## 5.5 Υλοποίηση σε Python

Το ακόλουθο script παρουσιάζει την ιδέα του "γραμμικού μοντέλου σε χαρακτηριστικά" σε ένα απλοποιημένο πρόβλημα μη-γραμμικής παλινδρόμησης. Δημιουργούμε μια μικρή λίστα δειγμάτων με θόρυβο από μια μη-γραμμική πραγματική συνάρτηση

$$y_{\text{true}}(x) = 0.6\sin(2\pi x) + 0.3x^3 - 0.2x,$$

και στη συνέχεια την προσαρμόζουμε χρησιμοποιώντας ένα μοντέλο που είναι *γραμμικό ως προς τις παραμέτρους* αλλά *μη-γραμμικό ως προς την είσοδο* μέσω μιας απεικόνισης χαρακτηριστικών

$$\boldsymbol{\phi}(x) = \left[1,\ x,\ x^2,\ x^3,\ \sin(2\pi x),\ \cos(2\pi x)\right]^{\mathsf{T}}.$$

Στοιβάζοντας αυτά τα χαρακτηριστικά για όλα τα δείγματα προκύπτει ένας πίνακας σχεδίασης $\boldsymbol{\Phi}$ και το μοντέλο γραμμικής παλινδρόμησης

$$\boldsymbol{y} \approx \boldsymbol{\Phi}\boldsymbol{\theta}.$$

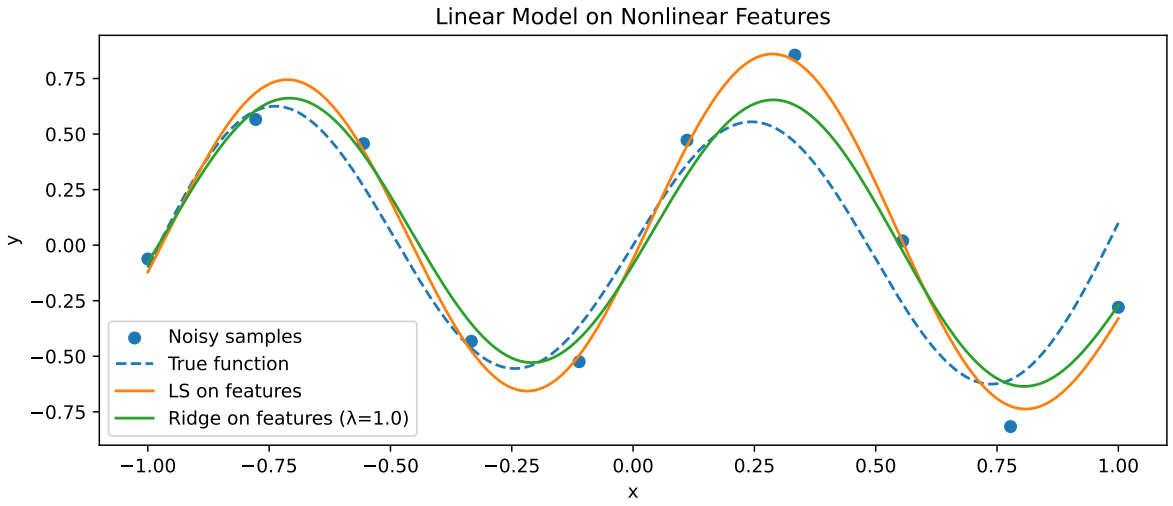
Συγκρίνονται δύο εκτιμητές:

- **Ελάχιστα Τετράγωνα (LS):** $\hat{\boldsymbol{\theta}}_{\text{LS}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{\theta}\|_2^2$ (υλοποίηση με `np.linalg.lstsq`).
- **Ridge (κανονικοποιημένα LS):** $\hat{\boldsymbol{\theta}}_{\text{ridge}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2$, που έχει κλειστή μορφή
$$\hat{\boldsymbol{\theta}}_{\text{ridge}} = (\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{y}.$$

Το Σχήμα 5.1 απεικονίζει τα θορυβώδη δείγματα, την πραγματική συνάρτηση και τις προσαρμοσμένες καμπύλες από LS και ridge. Το Σχήμα 5.2 απεικονίζει τα μαθημένα βάρη κάθε χαρακτηριστικού· η ridge συνήθως συρρικνώνει τους συντελεστές και μειώνει την ευαισθησία στον θόρυβο. Για αναφορά, το script κατασκευάζει επίσης το "πραγματικό" διάνυσμα συντελεστών (που αντιστοιχεί στη συνάρτηση που παρήγαγε τα δεδομένα σε αυτή τη βάση) και το συγκρίνει με τα εκτιμώμενα βάρη.

```
import numpy as np
import matplotlib.pyplot as plt

# ------------------------------
# 1) Generate nonlinear data
# ------------------------------
rng = np.random.default_rng(0)

N = 10
x = np.linspace(-1, 1, N)
y_true = 0.6*np.sin(2*np.pi*x) + 0.3*x**3 - 0.2*x
y = y_true + 0.3*rng.standard_normal(N)  # noisy observations

# Dense grid for smooth plotting
x_plot = np.linspace(-1, 1, 500)
y_true_plot = 0.6*np.sin(2*np.pi*x_plot) + 0.3*x_plot**3 - 0.2*x_plot

# ------------------------------
# 2) Feature map phi(x)
#    (polynomials + sin/cos)
# ------------------------------
def phi(x):
```

```
    """
    Feature vector:
    [1, x, x^2, x^3, sin(2πx), cos(2πx)]
    """
    x = np.asarray(x)
    return np.column_stack([
        np.ones_like(x),
        x,
        x**2,
        x**3,
        np.sin(2*np.pi*x),
        np.cos(2*np.pi*x),
    ])

Phi = phi(x)
Phi_plot = phi(x_plot)

# ------------------------------
# 3) Least Squares fit
#    y ≈ Phi theta
# ------------------------------
theta_ls, *_ = np.linalg.lstsq(Phi, y, rcond=None)
y_ls_plot = Phi_plot @ theta_ls

# ------------------------------
# 4) Ridge fit (regularized LS)
#    min ||y-Phi theta||^2 + λ||theta||^2
# ------------------------------
lam = 1.
I = np.eye(Phi.shape[1])
theta_ridge = np.linalg.solve(Phi.T @ Phi + lam*I, Phi.T @ y)
y_ridge_plot = Phi_plot @ theta_ridge

# ------------------------------
# 5) Plot results
# ------------------------------
plt.figure(figsize=(9, 4))

plt.scatter(x, y, label="Noisy samples", marker="o")
plt.plot(x_plot, y_true_plot, linestyle="--", label="True function")
plt.plot(x_plot, y_ls_plot, label="LS on features")
plt.plot(x_plot, y_ridge_plot, label=f"Ridge on features (λ={lam})")

plt.xlabel("x")
plt.ylabel("y")
plt.title("Linear Model on Nonlinear Features")
plt.legend()
plt.tight_layout()
plt.show()

# ------------------------------
# 6) Plot feature weights
# ------------------------------
feature_names = ["1", "x", "x^2", "x^3", "sin(2πx)", "cos(2πx)"]
theta_true = np.zeros_like(theta_ridge)
theta_true[1] = -0.2
theta_true[3] = 0.3
theta_true[4] = 0.6
```




Σχήμα 5.1: Θορυβώδη δείγματα, πραγματική συνάρτηση, προσαρμογή LS σε χαρακτηριστικά και προσαρμογή ridge με κανονικοποίηση.

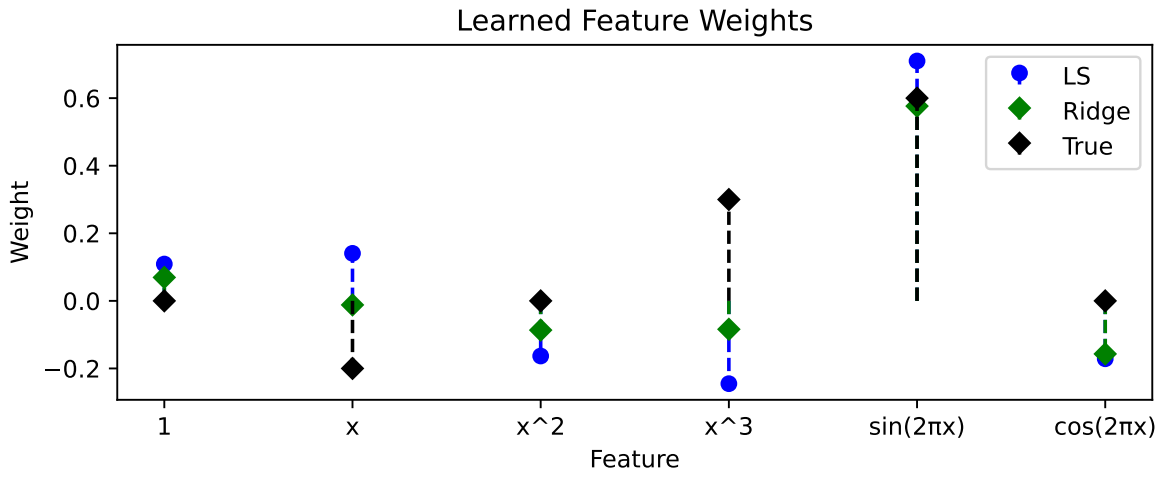
```
81
82 plt.figure(figsize=(7, 3))
83 idx = np.arange(len(feature_names))
84 plt.stem(idx, theta_ls, linefmt='b--', markerfmt='bo', basefmt=" ",
       label="LS")
85 plt.stem(idx, theta_ridge, linefmt="g--", markerfmt="gD", basefmt=" ",
       label="Ridge")
86 plt.stem(idx, theta_true, linefmt="k--", markerfmt="kD", basefmt=" ",
       label="True")
87
88 plt.xticks(idx, feature_names)
89 plt.xlabel("Feature")
90 plt.ylabel("Weight")
91 plt.title("Learned Feature Weights")
92 plt.legend()
93 plt.tight_layout()
94 plt.show()
95
96 print("theta_ls    =", theta_ls)
97 print("theta_ridge=", theta_ridge)
```

Listing 1: Παράδειγμα Γραμμικής Παλινδρόμησης σε Χαρακτηριστικά

# 6 Συμπερασματολογία κατά Bayes

## 6.1 Βασική Ιδέα

Στη συμπερασματολογία κατά Bayes, οι άγνωστες παράμετροι του μοντέλου αντιμετωπίζονται ως *τυχαίες μεταβλητές* και όχι ως σταθερές αλλά άγνωστες ποσότητες. Αντί να αναζητούμε μία μοναδική "βέλτιστη" εκτίμηση του $\boldsymbol{\theta}$, αναπαριστούμε την αβεβαιότητά μας σχετικά με την τιμή του μέσω μιας κατανομής πιθανότητας. Η αβεβαιότητα αυτή, πριν παρατηρήσουμε οποιαδήποτε δεδομένα, κωδικοποιείται μέσω μιας *πρότερης κατανομής* $p(\boldsymbol{\theta})$, η οποία αντικατοπτρίζει προϋπάρχουσα γνώση, φυσικούς περιορισμούς ή υποθέσεις μον-




Σχήμα 5.2: Εκτιμημένα βάρη χαρακτηριστικών για LS έναντι ridge.

τελοποίησης.

Τα παρατηρούμενα δεδομένα μοντελοποιούνται επίσης πιθανοκρατικά. Δοθέντων των εισόδων $\boldsymbol{X}$ και των παραμέτρων $\boldsymbol{\theta}$, υποθέτουμε ότι οι έξοδοι $\boldsymbol{Y}$ παράγονται σύμφωνα με ένα στοχαστικό μοντέλο δεδομένων. Αυτό λαμβάνει ρητά υπόψη τον θόρυβο μέτρησης, τη μη-μοντελοποιημένη δυναμική και τη στοχαστικότητα στη διαδικασία παραγωγής των δεδομένων.

Ο κεντρικός στόχος της Μπεϋζιανής συμπερασματολογίας, επομένως, δεν είναι να υπολογιστεί ένα μοναδικό διάνυσμα παραμέτρων, αλλά να εξαχθεί μια *κατανομή πάνω στις παραμέτρους* υπό συνθήκη στα παρατηρούμενα δεδομένα. Η κατανομή αυτή ποσοτικοποιεί τόσο το ποιες τιμές των παραμέτρων είναι πιθανές, όσο και το πόσο αβέβαιοι είμαστε για αυτές. Οι Μπεϋζιανές μέθοδοι είναι ιδιαίτερα ελκυστικές, διότι παρέχουν θεμελιωμένες εκτιμήσεις αβεβαιότητας, ενσωματώνουν την κανονικοποίηση φυσικά μέσω των πρότερων κατανομών, και επιτρέπουν συστηματική σύγκριση μεταξύ ανταγωνιστικών μοντέλων.

## 6.2 Πιθανοφάνεια

Η πιθανοφάνεια (likelihood)

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})$$

περιγράφει τον τρόπο με τον οποίο παράγονται τα δεδομένα δεδομένων των παραμέτρων. Πρόκειται για ένα πιθανοκρατικό μοντέλο των παρατηρήσεων υπό συνθήκη στο $\boldsymbol{\theta}$ (και ενδεχομένως στις εισόδους $\boldsymbol{X}$).

Η πιθανοφάνεια μπορεί να ερμηνευθεί με δύο συμπληρωματικούς τρόπους.

1. **Για σταθερές παραμέτρους $\boldsymbol{\theta}$**, είναι μια κατανομή πιθανότητας πάνω σε πιθανά σύνολα δεδομένων, περιγράφοντας πόσο πιθανές είναι διαφορετικές παρατηρήσεις υπό το θεωρούμενο μοντέλο.

2. **Για σταθερά παρατηρούμενα δεδομένα $(\boldsymbol{X}, \boldsymbol{Y})$**, γίνεται συνάρτηση του $\boldsymbol{\theta}$ που μετρά πόσο συμβατές είναι διαφορετικές τιμές παραμέτρων με τα δεδομένα.

Η επιλογή της πιθανοφάνειας εξαρτάται από το υποτιθέμενο μοντέλο θορύβου και τη διαδικασία παραγωγής δεδομένων.

Ένα συνηθισμένο παράδειγμα είναι ο προσθετικός Γκαουσιανός θόρυβος:

$$y_i = f(x_i; \boldsymbol{\theta}) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Υπό την υπόθεση ότι τα δείγματα θορύβου είναι ανεξάρτητα, η πιθανοφάνεια παραγοντοποιείται ως

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) = \prod_{i=1}^{N} \mathcal{N}\big(y_i \,;\, f(x_i; \boldsymbol{\theta}), \sigma^2\big)\,.$$

Αυτή η επιλογή συνδέει άμεσα τη Μπεϋζιανή συμπερασματολογία με την εκτίμηση ελαχίστων τετραγώνων, αφού η μεγιστοποίηση της πιθανοφάνειας υπό Γκαουσιανό θόρυβο είναι ισοδύναμη με την ελαχιστοποίηση ενός αθροίσματος τετραγώνων υπολοίπων.

## 6.3 Κανόνας του Bayes και η Εκ των Υστέρων Κατανομή

Ο κανόνας του Bayes συνδυάζει την πρότερη γνώση και τα δεδομένα μέσω

$$p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}) = \frac{p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\boldsymbol{Y} \mid \boldsymbol{X})}.$$

Η προκύπτουσα κατανομή $p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X})$ είναι η *εκ των υστέρων κατανομή* (posterior), η οποία αναπαριστά την ενημερωμένη πεποίθησή μας σχετικά με τις παραμέτρους μετά την παρατήρηση των δεδομένων.

Ο παρονομαστής

$$p(\boldsymbol{Y} \mid \boldsymbol{X}) = \int p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

ονομάζεται *τεκμήριο* (evidence) ή *περιθώρια πιθανοφάνεια* (marginal likelihood). Εξασφαλίζει ότι η εκ των υστέρων κατανομή ολοκληρώνεται στο ένα και παίζει βασικό ρόλο στη σύγκριση μοντέλων. Σημαντικό είναι ότι δεν εξαρτάται από το $\boldsymbol{\theta}$ όταν κάνουμε εκτίμηση παραμέτρων.

Μια χρήσιμη σύνοψη είναι

$$\text{posterior} \;\propto\; \text{likelihood} \times \text{prior}.$$

Η πιθανοφάνεια ωθεί την εκ των υστέρων κατανομή προς τιμές παραμέτρων που εξηγούν καλά τα δεδομένα, ενώ η πρότερη την ωθεί προς τιμές που θεωρούνται εύλογες πριν δούμε τα δεδομένα. Η εκ των υστέρων κατανομή αντιπροσωπεύει μια ισορροπία ανάμεσα σε αυτές τις δύο πηγές πληροφορίας.

## 6.4 Ερμηνεία των Μπεϋζιανών Ποσοτήτων

Κάθε συνιστώσα του κανόνα του Bayes έχει μια σαφή ερμηνεία. Η πρότερη κατανομή $p(\boldsymbol{\theta})$ κωδικοποιεί υποθέσεις ή γνώση πεδίου που είναι διαθέσιμες πριν παρατηρήσουμε δεδομένα. Η πιθανοφάνεια $p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})$ μετρά πόσο καλά μια συγκεκριμένη επιλογή παραμέτρων εξηγεί τα παρατηρούμενα δεδομένα. Η εκ των υστέρων κατανομή $p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X})$ είναι η ενημερωμένη πεποίθηση που συνδυάζει και τα δύο. Τέλος, η περιθώρια πιθανοφάνεια $p(\boldsymbol{Y} \mid \boldsymbol{X})$ ποσοτικοποιεί πόσο εύλογα είναι τα παρατηρούμενα δεδομένα υπό την υποτιθέμενη κλάση μοντέλων και την πρότερη κατανομή, και συνεπώς είναι χρήσιμη για τη σύγκριση διαφορετικών μοντέλων.

Από την οπτική της εκτίμησης, η Μπεϋζιανή συμπερασματολογία γενικεύει τις μεθόδους που μελετήθηκαν μέχρι τώρα: τα ελάχιστα τετράγωνα και η κανονικοποίηση εμφανίζονται ως ειδικές περιπτώσεις που αντιστοιχούν σε συγκεκριμένους συνδυασμούς πιθανοφάνειας–πρότερης κατανομής, ενώ οι Μπεϋζιανές μέθοδοι παρέχουν επιπλέον θεμελιωμένη ποσοτικοποίηση αβεβαιότητας και μια ενοποιημένη πιθανοκρατική ερμηνεία.

**Περιθώρια Πιθανοφάνεια και Σύγκριση Μοντέλων** Πέρα από την εκτίμηση παραμέτρων, η Μπεϋζιανή συμπερασματολογία παρέχει επίσης έναν θεμελιωμένο μηχανισμό για τη σύγκριση διαφορετικών μοντέλων. Αυτό γίνεται μέσω της *περιθώριας πιθανοφάνειας* (marginal likelihood),

$$p(\boldsymbol{Y} \mid \boldsymbol{X}) = \int p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Σε αντίθεση με την πιθανοφάνεια, η οποία αξιολογεί πόσο καλά μια συγκεκριμένη τιμή παραμέτρων εξηγεί τα δεδομένα, η περιθώρια πιθανοφάνεια μετρά πόσο καλά μια ολόκληρη κλάση μοντέλων εξηγεί τα δεδομένα *πριν* γνωρίσουμε τις παραμέτρους.

Η περιθώρια πιθανοφάνεια μπορεί να ερμηνευθεί ως ένας μέσος όρος της πιθανοφάνειας ως προς την πρότερη κατανομή. Μοντέλα που αποδίδουν υψηλή πιθανοφάνεια στα παρατηρούμενα δεδομένα για πολλές εύλογες τιμές παραμέτρων επιτυγχάνουν υψηλή περιθώρια πιθανοφάνεια, ενώ μοντέλα που απαιτούν πολύ «συγκεκριμένες» (fine-tuned) επιλογές παραμέτρων για να εξηγήσουν τα δεδομένα τιμωρούνται. Με αυτή την έννοια, η περιθώρια πιθανοφάνεια εξισορροπεί φυσικά την προσαρμογή στα δεδομένα και την πολυπλοκότητα του μοντέλου.

Αυτή η ποσότητα επιτρέπει τη Μπεϋζιανή *σύγκριση μοντέλων*. Δοθέντων δύο ανταγωνιστικών μοντέλων $\mathcal{M}_1$ και $\mathcal{M}_2$, η σχετική υποστήριξή τους από τα δεδομένα δίνεται από τον *παράγοντα Bayes* (Bayes factor):

$$\frac{p(\boldsymbol{Y} \mid \boldsymbol{X}, \mathcal{M}_1)}{p(\boldsymbol{Y} \mid \boldsymbol{X}, \mathcal{M}_2)}.$$

Αν αυτός ο λόγος είναι μεγαλύτερος από το ένα, τα δεδομένα ευνοούν το $\mathcal{M}_1$ έναντι του $\mathcal{M}_2$· αν είναι μικρότερος από το ένα, προτιμάται το $\mathcal{M}_2$.

Μια βασική εννοιολογική συνέπεια είναι το ενσωματωμένο αποτέλεσμα της θεωρίας του *ξυραφιού του Occam* (Occam's razor) στη Μπεϋζιανή συμπερασματολογία. Τα πιο σύνθετα μοντέλα τυπικά «απλώνουν» τη μάζα πιθανότητας σε έναν μεγαλύτερο χώρο παραμέτρων και επομένως τείνουν να έχουν μικρότερη περιθώρια πιθανοφάνεια, εκτός αν τα δεδομένα παρέχουν ισχυρή υποστήριξη για την επιπλέον πολυπλοκότητα. Ως αποτέλεσμα, η Μπεϋζιανή σύγκριση μοντέλων πραγματοποιεί αυτόματα έναν συμβιβασμό μεταξύ ποιότητας προσαρμογής και πολυπλοκότητας, χωρίς την ανάγκη για ad hoc ποινές ή ευριστικές διασταυρούμενης επικύρωσης (cross-validation).

## 6.5 Μπεϋζιανή Σημειακή Εκτίμηση μέσω Ελαχιστοποίησης Κινδύνου

Η Μπεϋζιανή συμπερασματολογία αποδίδει μια πλήρη εκ των υστέρων κατανομή $p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X})$, η οποία ποσοτικοποιεί την αβεβαιότητα σχετικά με τις άγνωστες παραμέτρους. Σε πολλές εφαρμογές, ωστόσο, επιθυμούμε τελικά να αναφέρουμε ένα μοναδικό διάνυσμα παραμέτρων $\hat{\boldsymbol{\theta}}$ για σκοπούς πρόβλεψης, ελέγχου ή ερμηνείας. Οι Μπεϋζιανοί σημειακοί εκτιμητές κατανοούνται καλύτερα μέσα από μια θεωρητική οπτική αποφάσεων: επιλέγουμε μια εκτίμηση που ελαχιστοποιεί την *αναμενόμενη εκ των υστέρων απώλεια* (posterior risk).

Δοθέντος μίας συνάρτησης κόστους (loss function) $\mathcal{L}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})$, ο βέλτιστος Μπεϋζιανός κανόνας απόφασης είναι

$$\hat{\boldsymbol{\theta}} = \arg\min_{\tilde{\boldsymbol{\theta}}} \mathbb{E}\Big[\mathcal{L}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) \;\Big|\; \boldsymbol{Y}, \boldsymbol{X}\Big] = \arg\min_{\tilde{\boldsymbol{\theta}}} \int \mathcal{L}(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})\, p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X})\, d\boldsymbol{\theta}.$$

Διαφορετικές συναρτήσεις κόστους αντιστοιχούν σε διαφορετικές έννοιες του «βέλτιστου» εκτιμητή και, συνεπώς, οδηγούν σε διαφορετικούς σημειακούς εκτιμητές.

### 6.5.1 Εκτιμητής MAP (Τρόπος της Εκ των Υστέρων Κατανομής)

Ο εκτιμητής μέγιστης εκ των υστέρων πιθανότητας (Maximum A Posteriori, MAP) επιλέγει την τιμή της παραμέτρου στην οποία μεγιστοποιείται η εκ των υστέρων πυκνότητα:

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}).$$

Χρησιμοποιώντας τον κανόνα του Bayes,

$$p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}) \propto p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})\, p(\boldsymbol{\theta}),$$

οπότε ισοδύναμα

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})\, p(\boldsymbol{\theta}).$$

Σε μορφή βελτιστοποίησης, ο MAP προκύπτει ελαχιστοποιώντας το αρνητικό λογάριθμο της εκ των υστέρων κατανομής:

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \arg\min_{\boldsymbol{\theta}} \Big( -\log p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) - \log p(\boldsymbol{\theta}) \Big).$$

Έτσι, η εκτίμηση MAP μπορεί να ερμηνευθεί ως η ελαχιστοποίηση ενός *όρου ασυμφωνίας με τα δεδομένα* (που προέρχεται από την πιθανοφάνεια) συν ενός *όρου κανονικοποίησης* (που προέρχεται από την πρότερη κατανομή). Επειδή ο MAP επιλέγει την επικρατούσα τιμή (mode) της εκ των υστέρων κατανομής, επιλέγει τη μοναδική πιο πιθανή τιμή παραμέτρων (αν η εκ των υστέρων κατανομή είναι πολυτροπική, ενδέχεται να υπάρχουν πολλαπλές λύσεις MAP — πολλαπλά τοπικά μέγιστα).

**MAP, Πρότερες Κατανομές και Κανονικοποίηση** Η σύνδεση μεταξύ προτέρων κατανομών και κανονικοποίησης γίνεται ρητή σε συνηθισμένους συνδυασμούς πιθανοφάνειας–πρότερης κατανομής. Για παράδειγμα, μια ισοτροπική Γκαουσιανή πρότερη κατανομή μηδενικού μέσου

$$p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \alpha^{-1}\boldsymbol{I})$$

συνεπάγεται

$$-\log p(\boldsymbol{\theta}) = \text{σταθ.} + \frac{\alpha}{2}\|\boldsymbol{\theta}\|_2^2,$$

δηλαδή έναν όρο ποινής $\ell_2$ στις παραμέτρους. Αν η πιθανοφάνεια αντιστοιχεί σε Γκαουσιανό θόρυβο παρατήρησης,

$$y_i = f(x_i; \boldsymbol{\theta}) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

τότε

$$-\log p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) = \text{σταθ.} + \frac{1}{2\sigma^2}\|\boldsymbol{Y} - f(\boldsymbol{X}; \boldsymbol{\theta})\|_2^2.$$

Επομένως, ο MAP ανάγεται σε ένα πρόβλημα κανονικοποιημένων ελαχίστων τετραγώνων:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma^2}\|\boldsymbol{Y} - f(\boldsymbol{X};\boldsymbol{\theta})\|_2^2 + \frac{\alpha}{2}\|\boldsymbol{\theta}\|_2^2.$$

Αυτό παρέχει μια ακριβή πιθανοκρατική ερμηνεία της κανονικοποίησης ridge (Tikhonov): αντιστοιχεί σε μια Γκαουσιανή πρότερη κατανομή στις παραμέτρους.

**MLE ως Ειδική Περίπτωση του MAP** Αν η πρότερη κατανομή είναι επίπεδη στην περιοχή ενδιαφέροντος των παραμέτρων,

$$p(\boldsymbol{\theta}) \propto \text{σταθ.},$$

τότε ο εκτιμητής MAP ανάγεται στον εκτιμητή μέγιστης πιθανοφάνειας (Maximum Likelihood Estimator, MLE):

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) = \hat{\boldsymbol{\theta}}_{\text{MLE}}.$$

Με λόγια, ο MAP αξιοποιεί τόσο την πρότερη πληροφορία όσο και τα δεδομένα, ενώ ο MLE χρησιμοποιεί μόνο τα δεδομένα. Καθώς η ποσότητα των δεδομένων αυξάνεται, η επίδραση της πρότερης κατανομής συνήθως μειώνεται και οι MAP και MLE συχνά συγκλίνουν σε παρόμοιες λύσεις.

### 6.5.2 Εκτιμητής MMSE (εκτιμητής ελάχιστου μέσου τετραγωνικού σφάλματος)

Ο εκτιμητής ελάχιστου μέσου τετραγωνικού σφάλματος (Minimum Mean Square Error, MMSE) προκύπτει από τη χρήση τετραγωνικής συνάρτησης κόστους:

$$\hat{\boldsymbol{\theta}}_{\text{MMSE}} = \arg\min_{\tilde{\boldsymbol{\theta}}} \mathbb{E}\Big[\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2^2 \;\Big|\; \boldsymbol{Y}, \boldsymbol{X}\Big].$$

Ο μοναδικός ελαχιστοποιητής είναι ο μέσος της εκ των υστέρων κατανομής:

$$\hat{\boldsymbol{\theta}}_{\text{MMSE}} = \mathbb{E}[\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}] = \int \boldsymbol{\theta}\, p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X})\, d\boldsymbol{\theta}.$$

Διαισθητικά, ο MMSE υπολογίζει έναν μέσο όρο όλων των εύλογων τιμών των παραμέτρων, σταθμισμένο από την εκ των υστέρων πιθανότητά τους. Αυτό μπορεί να είναι πιο ευρωστό από τον MAP όταν η εκ των υστέρων κατανομή είναι ασύμμετρη ή πολυτροπική, διότι ο μέσος λαμβάνει υπόψη ολόκληρη τη μάζα της κατανομής και όχι μόνο την κορυφή της.

**Ανάλυση.** Έστω $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}]$. Θεωρούμε τον εκ των υστέρων κίνδυνο για μια υποψήφια εκτίμηση $\tilde{\boldsymbol{\theta}}$:

$$\mathbb{E}\Big[\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2^2 \mid \boldsymbol{Y}, \boldsymbol{X}\Big].$$

Προσθέτουμε και αφαιρούμε το $\boldsymbol{\mu}$:

$$\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2^2 = \|(\boldsymbol{\theta} - \boldsymbol{\mu}) + (\boldsymbol{\mu} - \tilde{\boldsymbol{\theta}})\|_2^2 = \|\boldsymbol{\theta} - \boldsymbol{\mu}\|_2^2 + \|\boldsymbol{\mu} - \tilde{\boldsymbol{\theta}}\|_2^2 + 2(\boldsymbol{\theta} - \boldsymbol{\mu})^{\mathsf{T}}(\boldsymbol{\mu} - \tilde{\boldsymbol{\theta}}).$$

Λαμβάνοντας την εκ των υστέρων αναμενόμενη τιμή, ο διασταυρούμενος όρος μηδενίζεται επειδή $\mathbb{E}[\boldsymbol{\theta} - \boldsymbol{\mu} \mid \boldsymbol{Y}, \boldsymbol{X}] = \mathbf{0}$. Συνεπώς,

$$\mathbb{E}\Big[\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2^2 \mid \cdot\Big] = \underbrace{\mathbb{E}\big[\|\boldsymbol{\theta} - \boldsymbol{\mu}\|_2^2 \mid \cdot\big]}_{\text{ανεξάρτητο του } \tilde{\boldsymbol{\theta}}} + \|\boldsymbol{\mu} - \tilde{\boldsymbol{\theta}}\|_2^2,$$

το οποίο ελαχιστοποιείται όταν $\tilde{\boldsymbol{\theta}} = \boldsymbol{\mu}$.

### 6.5.3 Εκτιμητής MMAE (Διάμεσος της Εκ των Υστέρων Κατανομής)

Αν, αντί για τετραγωνική, χρησιμοποιήσουμε απόλυτη συνάρτηση κόστους (κατά συνιστώσα απώλεια $\ell_1$),

$$\hat{\boldsymbol{\theta}}_{\text{MMAE}} = \arg\min_{\tilde{\boldsymbol{\theta}}} \mathbb{E}\Big[\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_1 \;\Big|\; \boldsymbol{Y}, \boldsymbol{X}\Big], \qquad \|\boldsymbol{v}\|_1 = \sum_i |v_i|,$$

τότε ο βέλτιστος εκτιμητής είναι η εκ των υστέρων διάμεσος. Στην πράξη, μια συνηθισμένη επιλογή είναι η *κατά συνιστώσα εκ των υστέρων διάμεσος*:

$$\hat{\boldsymbol{\theta}}_{\text{MMAE}} = \big[\text{διάμεσος}(p(\theta_1 \mid \boldsymbol{Y}, \boldsymbol{X})), \ldots, \text{διάμεσος}(p(\theta_P \mid \boldsymbol{Y}, \boldsymbol{X}))\big]^{\mathsf{T}}.$$

Ο MMAE είναι συχνά πιο ευρωστός σε κατανομές με «βαριές» ουρές (heavy-tailed) και σε ακραίες τιμές (outliers) από τον MMSE, διότι οι διάμεσοι επηρεάζονται λιγότερο από ακραίες τιμές σε σχέση με τους μέσους.

### 6.5.4 Σύνοψη: MAP vs. MMSE vs. MMAE

Οι τρεις πιο συνηθισμένοι Μπεϋζιανοί σημειακοί εκτιμητές αντιστοιχούν σε διαφορετικές περιλήψεις της εκ των υστέρων κατανομής:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}) \quad \text{(mode)}, \qquad \hat{\boldsymbol{\theta}}_{\text{MMSE}} = \mathbb{E}[\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}] \quad \text{(μέσος)},$$

$$\hat{\boldsymbol{\theta}}_{\text{MMAE}} = \arg\min_{\tilde{\boldsymbol{\theta}}} \mathbb{E}\Big[\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_1 \mid \boldsymbol{Y}, \boldsymbol{X}\Big] \quad \text{(διάμεσος)}.$$

Αν η εκ των υστέρων κατανομή είναι Γκαουσιανή (ή, γενικότερα, συμμετρική και μονοτροπική), τότε

$$\text{μέσος} = \text{διάμεσος} = \text{mode} \quad \Rightarrow \quad \hat{\boldsymbol{\theta}}_{\text{MMSE}} = \hat{\boldsymbol{\theta}}_{\text{MMAE}} = \hat{\boldsymbol{\theta}}_{\text{MAP}}.$$

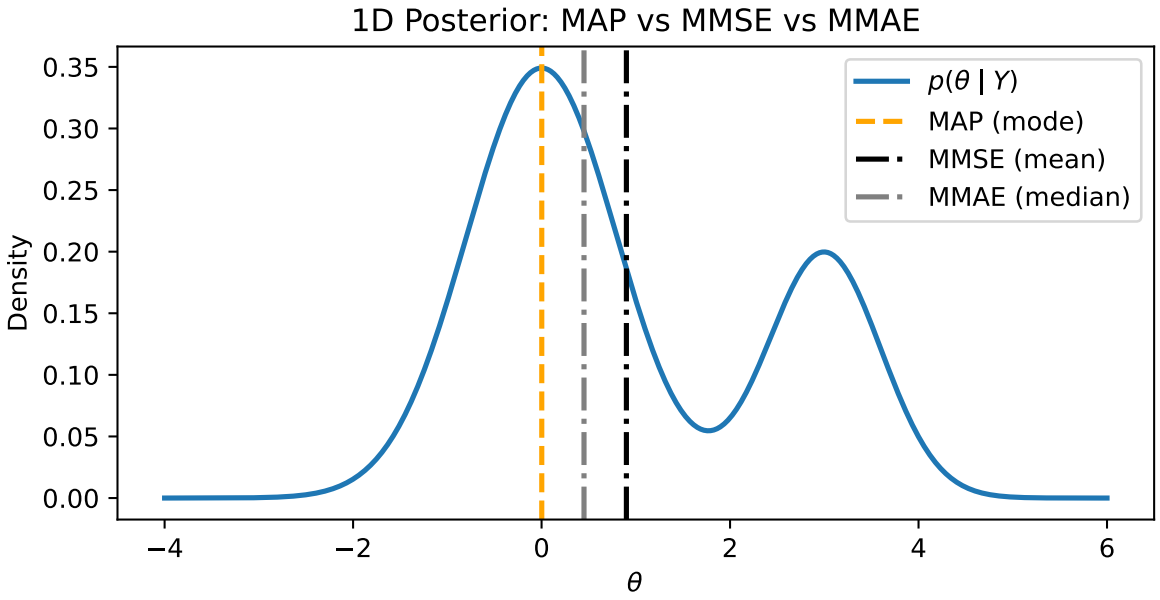
Για ασύμμετρες, πολυτροπικές ή με βαριές ουρές εκ των υστέρων κατανομές, οι εκτιμητές αυτοί μπορεί να διαφέρουν σημαντικά, και η κατάλληλη επιλογή εξαρτάται από τη συνάρτηση κόστους που αποτυπώνει καλύτερα την έννοια του σφάλματος στην εκάστοτε εφαρμογή.

**Γεωμετρική Διαίσθηση** Το Σχήμα 6.1 απεικονίζει τη γεωμετρική διαίσθηση πίσω από τους MAP, MMSE και MMAE στη μία διάσταση. Ο εκτιμητής MAP αντιστοιχεί στην κορυφή της εκ των υστέρων πυκνότητας (mode), ο εκτιμητής MMSE στο κέντρο μάζας (μέσος), και ο εκτιμητής MMAE στο σημείο που χωρίζει την εκ των υστέρων κατανομή σε δύο ίσα μέρη πιθανότητας (διάμεσος). Το βασικό συμπέρασμα είναι ότι διαφορετικοί εκτιμητές αντανακλούν διαφορετικά κριτήρια βελτιστότητας (διαφορετικές συναρτήσεις κόστους).

## 6.6 Εκτίμηση Μέγιστης Πιθανοφάνειας (Maximum Likelihood Estimation, MLE)

Η εκτίμηση μέγιστης πιθανοφάνειας είναι μια συχνοτική (frequentist) προσέγγιση, στην οποία το άγνωστο διάνυσμα παραμέτρων $\boldsymbol{\theta}$ αντιμετωπίζεται ως μια σταθερή (αλλά άγνωστη) ποσότητα. Υποθέτουμε ένα μοντέλο παρατήρησης που καθορίζει μια πιθανοφάνεια

$$p(\boldsymbol{y} \mid \boldsymbol{\theta}),$$




Σχήμα 6.1: Γεωμετρική διαίσθηση για Μπεϋζιανούς σημειακούς εκτιμητές στη μία διάσταση. Για μια εκ των υστέρων κατανομή $p(\theta \mid Y)$: ο εκτιμητής MAP $\hat{\theta}_{\mathrm{MAP}}$ είναι η *επικρατούσα τιμή (mode)* της κατανομής, ο εκτιμητής MMSE $\hat{\theta}_{\mathrm{MMSE}}$ είναι ο *μέσος* (κέντρο μάζας), και ο εκτιμητής MMAE $\hat{\theta}_{\mathrm{MMAE}}$ είναι η *διάμεσος* (50% μάζα πιθανότητας σε κάθε πλευρά). Διαφορετικοί εκτιμητές αντιστοιχούν σε διαφορετικές συναρτήσεις κόστους και μπορεί να διαφέρουν σημαντικά για ασύμμετρες ή πολυτροπικές εκ των υστέρων κατανομές.

δηλαδή την πιθανότητα (ή πυκνότητα πιθανότητας) να παρατηρηθούν τα δεδομένα $\boldsymbol{y}$ για μια δεδομένη τιμή των παραμέτρων.

Ο MLE επιλέγει την τιμή των παραμέτρων που καθιστά τα παρατηρούμενα δεδομένα όσο το δυνατόν πιο πιθανά:

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\theta}} \; p(\boldsymbol{y} \mid \boldsymbol{\theta}).$$

Στην πράξη, σχεδόν πάντα εργαζόμαστε σε λογαριθμική κλίμακα (log-likelihood)

$$\mathcal{L}(\boldsymbol{\theta}) = \log p(\boldsymbol{y} \mid \boldsymbol{\theta}), \qquad \hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\theta}} \; \mathcal{L}(\boldsymbol{\theta}),$$

διότι ο λογάριθμος μετατρέπει γινόμενα σε αθροίσματα και βελτιώνει την αριθμητική σταθερότητα, χωρίς να αλλάζει τον μεγιστοποιητή.

**Ανεξάρτητα και ομοιόμορφα κατανεμημένα δεδομένα (i.i.d.).** Αν τα δείγματα είναι ανεξάρτητα και ομοιόμορφα κατανεμημένα (independent and identically distributed, i.i.d.), τότε η πιθανοφάνεια παραγοντοποιείται:

$$p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{k=1}^{N} p(y_k \mid \boldsymbol{\theta}),$$

οπότε η λογαριθμική πιθανοφάνεια γίνεται άθροισμα:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{k=1}^{N} \log p(y_k \mid \boldsymbol{\theta}).$$

Πιο γενικά, για i.i.d. ζεύγη εισόδου–εξόδου $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ με ένα υπό συνθήκη μοντέλο,

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) = \prod_{i=1}^{N} p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}), \qquad \mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}),$$

και ο MLE είναι

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}).$$

Το κρίσιμο σημείο είναι ότι ο MLE εξαρτάται αποκλειστικά από το υποτιθέμενο μοντέλο δεδομένων/θορύβου μέσω της $p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$: αλλάζοντας την υπόθεση για το μοντέλο κατανομής, αλλάζει και ο εκτιμητής.

**Πώς υπολογίζεται ο MLE;** Αν η λογαριθμική πιθανοφάνεια είναι κοίλη ως προς $\boldsymbol{\theta}$ και έχει απλή μορφή, ο μεγιστοποιητής μπορεί να είναι διαθέσιμος σε κλειστή μορφή (ή μέσω μιας μοναδικής επίλυσης γραμμικού συστήματος). Διαφορετικά, επιλύουμε αριθμητικά ένα πρόβλημα βελτιστοποίησης, συνήθως χρησιμοποιώντας τα εργαλεία που έχουμε ήδη δει στη βελτιστοποίηση: κατάβαση/ανάβαση κλίσης στο $-\mathcal{L}$, τη μέθοδο του Newton, τη μέθοδο Gauss–Newton (όταν το αντικειμενικό έχει δομή ελαχίστων τετραγώνων), καθώς και στρατηγικές line search ή trust region για αυξημένη ευρωστία.

**Παράδειγμα: γραμμικό μοντέλο με Γκαουσιανό θόρυβο ⇒ ελάχιστα τετράγωνα.** Θεωρούμε το μοντέλο γραμμικής παλινδρόμησης

$$y_i = \boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{x}_i + w_i, \qquad \boldsymbol{\theta} \in \mathbb{R}^p,$$

με i.i.d. Γκαουσιανό θόρυβο $w_i \sim \mathcal{N}(0, \sigma^2)$. Τότε

$$p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) = \mathcal{N}\big(y_i; \boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{x}_i, \sigma^2\big)\,,$$

και η λογαριθμική πιθανοφάνεια είναι

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left[ -\frac{1}{2} \log\big(2\pi\sigma^2\big) - \frac{1}{2\sigma^2} \big(y_i - \boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{x}_i\big)^2 \right].$$

Η μεγιστοποίηση της $\mathcal{L}(\boldsymbol{\theta})$ είναι ισοδύναμη με την ελαχιστοποίηση του αθροίσματος τετραγώνων των υπολοίπων:

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) \iff \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} \big(y_i - \boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{x}_i\big)^2.$$

Επομένως, υπό γραμμικό μοντέλο με i.i.d. Γκαουσιανό θόρυβο, ο MLE ταυτίζεται ακριβώς με τα κλασικά γραμμικά ελάχιστα τετράγωνα. Αυτό παρέχει μια πιθανοκρατική αιτιολόγηση του κριτηρίου LS και αποσαφηνίζει πώς η υπόθεση για την κατανομή του θορύβου καθορίζει το αντικειμενικό της εκτίμησης.

### 6.6.1 Λυμένο Παράδειγμα: MLE για τη Μέση Τιμή Γκαουσιανής (Γνωστή Διακύμανση)

Έστω ότι παρατηρούμε i.i.d. δείγματα

$$y_1, \ldots, y_N \quad \text{με} \quad y_k \sim \mathcal{N}(\mu, \sigma^2),$$

όπου η διακύμανση $\sigma^2$ είναι γνωστή και η άγνωστη παράμετρος είναι η μέση τιμή $\theta = \mu$.

**Πιθανοφάνεια.** Εφόσον τα δείγματα είναι i.i.d.,

$$p(\boldsymbol{y} \mid \mu) = \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_k-\mu)^2}{2\sigma^2}\right).$$

**Λογαριθμική πιθανοφάνεια.** Λαμβάνοντας λογάριθμο,

$$\mathcal{L}(\mu) = \log p(\boldsymbol{y} \mid \mu) = -\frac{N}{2}\log\bigl(2\pi\sigma^2\bigr) - \frac{1}{2\sigma^2}\sum_{k=1}^{N}(y_k-\mu)^2.$$

Επειδή ο πρώτος όρος είναι σταθερός ως προς $\mu$, η μεγιστοποίηση της $\mathcal{L}(\mu)$ είναι ισοδύναμη με την ελαχιστοποίηση του $\sum_{k=1}^{N}(y_k-\mu)^2$.

**Υπολογισμός του μεγιστοποιητή.** Υπολογίζουμε την παράγωγο της $\mathcal{L}(\mu)$ (ή του ισοδύναμου αντικειμενικού αθροίσματος τετραγώνων) και την εξισώνουμε με το μηδέν:

$$\frac{d\mathcal{L}}{d\mu} = -\frac{1}{2\sigma^2}\frac{d}{d\mu}\sum_{k=1}^{N}(y_k-\mu)^2 = -\frac{1}{2\sigma^2}\sum_{k=1}^{N}2(y_k-\mu)(-1) = \frac{1}{\sigma^2}\sum_{k=1}^{N}(y_k-\mu).$$

Θέτουμε $\frac{d\mathcal{L}}{d\mu} = 0$:

$$\sum_{k=1}^{N}(y_k-\mu) = 0 \quad \Rightarrow \quad N\mu = \sum_{k=1}^{N} y_k \quad \Rightarrow \quad \boxed{\hat{\mu}_{\mathrm{MLE}} = \frac{1}{N}\sum_{k=1}^{N} y_k}.$$

Άρα, ο MLE της μέσης τιμής μιας Γκαουσιανής (με γνωστή διακύμανση) είναι ο δειγματικός μέσος.

**Αριθμητικό mini-example.** Έστω

$$y_1 = 1.2, \quad y_2 = 0.7, \quad y_3 = 1.5, \quad y_4 = 0.6, \quad y_5 = 1.0, \qquad (N = 5).$$

Τότε

$$\hat{\mu}_{\mathrm{MLE}} = \frac{1}{5}(1.2 + 0.7 + 1.5 + 0.6 + 1.0) = \frac{1}{5}\cdot 5.0 = 1.0.$$

**Σύνδεση με ελάχιστα τετράγωνα.** Σε αυτό το παράδειγμα, η μεγιστοποίηση της Γκαουσιανής πιθανοφάνειας είναι ισοδύναμη με την ελαχιστοποίηση του $\sum_{k=1}^{N}(y_k-\mu)^2$, δηλαδή με την προσαρμογή ενός σταθερού μοντέλου στα δεδομένα με την έννοια των ελαχίστων τετραγώνων. Αυτό αντικατοπτρίζει το γενικό γεγονός ότι ο MLE με Γκαουσιανό θόρυβο οδηγεί σε αντικειμενικά τετραγωνικού σφάλματος.

### 6.6.2 Λυμένο Παράδειγμα: MLE για Γραμμική Παλινδρόμηση με Άγνωστη Διακύμανση

Θεωρούμε το τυπικό γραμμικό μοντέλο με Γκαουσιανό θόρυβο:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{w}, \qquad \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}),$$

όπου τόσο οι συντελεστές παλινδρόμησης $\boldsymbol{\theta} \in \mathbb{R}^p$ όσο και η διακύμανση θορύβου $\sigma^2 > 0$ είναι άγνωστα.

**Πιθανοφάνεια.** Η υπό συνθήκη πυκνότητα του $\boldsymbol{y}$ δεδομένων των $(\boldsymbol{\theta}, \sigma^2)$ είναι πολυδιάστατη Γκαουσιανή:

$$p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\theta}, \sigma^2) = (2\pi\sigma^2)^{-N/2} \exp\left( -\frac{1}{2\sigma^2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 \right).$$

**Λογαριθμική πιθανοφάνεια.** Λαμβάνοντας λογάριθμο:

$$\mathcal{L}(\boldsymbol{\theta}, \sigma^2) = -\frac{N}{2} \log\bigl(2\pi\sigma^2\bigr) - \frac{1}{2\sigma^2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2.$$

**MLE για το $\boldsymbol{\theta}$.** Για σταθερό $\sigma^2$, η μεγιστοποίηση της $\mathcal{L}$ ως προς $\boldsymbol{\theta}$ είναι ισοδύναμη με την ελαχιστοποίηση του $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2$, άρα ο MLE για το $\boldsymbol{\theta}$ είναι η λύση κλασικών ελαχίστων τετραγώνων:

$$\boxed{\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}} \quad \text{(υποθέτοντας ότι ο } \boldsymbol{X} \text{ έχει πλήρη τάξη στηλών).}$$

**MLE για το $\sigma^2$.** Αντικαθιστούμε το $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ στη λογαριθμική πιθανοφάνεια και μεγιστοποιούμε ως προς $\sigma^2$. Παραγωγίζουμε την $\mathcal{L}$ ως προς $\sigma^2$:

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = -\frac{N}{2}\frac{1}{\sigma^2} + \frac{1}{2}\frac{1}{(\sigma^2)^2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2.$$

Θέτοντας το ίσο με μηδέν παίρνουμε

$$-N\sigma^2 + \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 = 0 \quad \Rightarrow \quad \sigma^2 = \frac{1}{N} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2.$$

Επομένως,

$$\boxed{\hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{N} \bigl\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}\bigr\|_2^2}$$

(σημείωση: ο αμερόληπτος εκτιμητής διακύμανσης χρησιμοποιεί $\frac{1}{N-p}$ αντί για $\frac{1}{N}$).

**Αριθμητικό παράδειγμα.** Προσαρμόζουμε μια ευθεία $y = \theta_0 + \theta_1 x + w$ από τρία δείγματα:

$$(x_1, y_1) = (0, 1), \quad (x_2, y_2) = (1, 2), \quad (x_3, y_3) = (2, 2).$$

Τότε

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}.$$

Υπολογίζουμε

$$\boldsymbol{X}^\mathsf{T}\boldsymbol{X} = \begin{bmatrix} 3 & 3 \\ 3 & 5 \end{bmatrix}, \qquad \boldsymbol{X}^\mathsf{T}\boldsymbol{y} = \begin{bmatrix} 5 \\ 6 \end{bmatrix}, \qquad (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1} = \frac{1}{6}\begin{bmatrix} 5 & -3 \\ -3 & 3 \end{bmatrix}.$$

Άρα

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \frac{1}{6}\begin{bmatrix} 5 & -3 \\ -3 & 3 \end{bmatrix}\begin{bmatrix} 5 \\ 6 \end{bmatrix} = \begin{bmatrix} 7/6 \\ 1/2 \end{bmatrix}, \quad \Rightarrow \quad \hat{y}(x) = \frac{7}{6} + \frac{1}{2}x.$$

Υπόλοιπα:

$$\hat{\boldsymbol{r}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \begin{bmatrix} 1 - 7/6 \\ 2 - (7/6 + 1/2) \\ 2 - (7/6 + 1) \end{bmatrix} = \begin{bmatrix} -1/6 \\ 1/3 \\ -1/6 \end{bmatrix}, \qquad \|\hat{\boldsymbol{r}}\|_2^2 = \frac{1}{36} + \frac{1}{9} + \frac{1}{36} = \frac{1}{6}.$$

Επομένως, με $N = 3$,

$$\hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{N}\|\hat{\boldsymbol{r}}\|_2^2 = \frac{1}{3} \cdot \frac{1}{6} = \frac{1}{18}.$$

# 7 Μπεϋζιανή Γραμμική Παλινδρόμηση

## 7.1 Μοντέλο, Συμβολισμός και Στόχοι

Η Μπεϋζιανή γραμμική παλινδρόμηση (Bayesian Linear Regression, BLR) συνδυάζει ένα μοντέλο *γραμμικό ως προς τις παραμέτρους* με μια πιθανοκρατική περιγραφή της αβεβαιότητας στις παραμέτρους. Παρατηρούμε δεδομένα εκπαίδευσης

$$\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N,$$

επιλέγουμε μια απεικόνιση χαρακτηριστικών

$$\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^P, \qquad \boldsymbol{\phi}_i \equiv \boldsymbol{\phi}(\boldsymbol{x}_i),$$

και υποθέτουμε το μοντέλο παρατήρησης

$$y_i = \boldsymbol{\phi}_i^{\mathsf{T}}\boldsymbol{\theta} + w_i, \qquad w_i \sim \mathcal{N}(0, \sigma_w^2), \quad \text{i.i.d.}$$

Στοιβάζοντας τις παρατηρήσεις προκύπτει

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \qquad \boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\phi}_1^{\mathsf{T}} \\ \vdots \\ \boldsymbol{\phi}_N^{\mathsf{T}} \end{bmatrix} \in \mathbb{R}^{N \times P}, \qquad \boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\theta} + \boldsymbol{w}, \quad \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \sigma_w^2 \boldsymbol{I}).$$

Ο Μπεϋζιανός στόχος είναι διττός:

- να εξαχθεί μια εκ των υστέρων κατανομή πάνω στα βάρη, $p(\boldsymbol{\theta} \mid \mathcal{D})$,
- να πραγματοποιηθεί πρόβλεψη σε ένα νέο σημείο εισόδου $\boldsymbol{x}_*$ με ποσοτικοποίηση αβεβαιότητας: $p(y_* \mid \boldsymbol{x}_*, \mathcal{D})$.

## 7.2 Πρότερη Κατανομή και Πιθανοφάνεια

**Πρότερη κατανομή στα βάρη (συζυγής Γκαουσιανή).** Κωδικοποιούμε την αβεβαιότητα και την εκ των προτέρων γνώση για τα βάρη μέσω μιας Γκαουσιανής πρότερης κατανομής

$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0).$$

Εδώ, το $\boldsymbol{\mu}_0$ είναι ο πρότερος μέσος όρος (η καλύτερη εκ των προτέρων εκτίμησή μας για τα βάρη πριν την παρατήρηση δεδομένων), ενώ το $\boldsymbol{\Sigma}_0$ είναι η πρότερη συνδιακύμανση (ποσοτικοποιεί την αβεβαιότητα και τις συσχετίσεις). Μια συχνή ειδική περίπτωση είναι η ισότροπη πρότερη κατανομή $\boldsymbol{\Sigma}_0 = \alpha^{-1}\boldsymbol{I}$, η οποία συρρικνώνει τα βάρη προς το $\boldsymbol{\mu}_0$ (συχνά $\boldsymbol{0}$).

**Πιθανοφάνεια.** Υπό συνθήκη στο $\boldsymbol{\theta}$, οι έξοδοι είναι Γκαουσιανές:

$$p(\boldsymbol{y} \mid \boldsymbol{\Phi}, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\Phi}\boldsymbol{\theta}, \sigma_w^2 \boldsymbol{I}).$$

Ισοδύναμα, η λογαριθμική πιθανοφάνεια είναι

$$\log p(\boldsymbol{y} \mid \boldsymbol{\Phi}, \boldsymbol{\theta}) = -\frac{N}{2}\log\bigl(2\pi\sigma_w^2\bigr) - \frac{1}{2\sigma_w^2}\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\|_2^2.$$

Έτσι, ο Γκαουσιανός θόρυβος συνδέεται άμεσα με τα ελάχιστα τετράγωνα (MLE $\Leftrightarrow$ LS).

## 7.3 Εκ των Υστέρων Κατανομή των Βαρών

Ο κανόνας του Bayes δίνει

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{\Phi}) \propto p(\boldsymbol{y} \mid \boldsymbol{\Phi}, \boldsymbol{\theta})\, p(\boldsymbol{\theta}).$$

Θα υπολογίσουμε ρητά την εκ των υστέρων κατανομή και θα δείξουμε ότι είναι Γκαουσιανή (συζυγία).

**Βήμα 1: Μη κανονικοποιημένος λογάριθμος της εκ των υστέρων κατανομής.** Αγνοώντας σταθερούς όρους ανεξάρτητους του $\boldsymbol{\theta}$,

$$\begin{aligned}\log p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{\Phi}) &= \log p(\boldsymbol{y} \mid \boldsymbol{\Phi}, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) + \text{const} \\ &= -\frac{1}{2\sigma_w^2}\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\|_2^2 - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_0)^\mathsf{T}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_0) + \text{const}.\end{aligned}$$

**Βήμα 2: Ανάπτυξη των τετραγωνικών μορφών.** Πρώτα,

$$\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\|_2^2 = \boldsymbol{y}^\mathsf{T}\boldsymbol{y} - 2\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{y} + \boldsymbol{\theta}^\mathsf{T}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\boldsymbol{\theta}.$$

Δεύτερον,

$$(\boldsymbol{\theta} - \boldsymbol{\mu}_0)^\mathsf{T}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_0) = \boldsymbol{\theta}^\mathsf{T}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\mu}_0^\mathsf{T}\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0.$$

Αντικαθιστώντας και συλλέγοντας όρους ως προς $\boldsymbol{\theta}$, παίρνουμε

$$\log p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{\Phi}) = -\frac{1}{2}\left[\boldsymbol{\theta}^\mathsf{T}\left(\boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma_w^2}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}\right)\boldsymbol{\theta} - 2\boldsymbol{\theta}^\mathsf{T}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \frac{1}{\sigma_w^2}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{y}\right)\right] + \text{const}.$$

**Βήμα 3: Αναγνώριση της Γκαουσιανής (συμπλήρωση τετραγώνου).** Ορίζουμε τη *posterior ακρίβεια* και τον αντίστοιχο γραμμικό όρο:

$$\boldsymbol{\Lambda}_N \triangleq \boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma_w^2}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}, \qquad \boldsymbol{\eta}_N \triangleq \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \frac{1}{\sigma_w^2}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{y}.$$

Τότε

$$\boldsymbol{\mu}_N = \boldsymbol{\Lambda}_N^{-1}\boldsymbol{\eta}_N, \qquad \boldsymbol{\Sigma}_N = \boldsymbol{\Lambda}_N^{-1},$$

και τελικά

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{\Phi}) = \mathcal{N}(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N).$$

**Ερμηνεία μέσω "πρόσθεσης ακριβείας".**

$$\boldsymbol{\Sigma}_N^{-1} = \boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma_w^2}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi}.$$

Τα δεδομένα προσθέτουν ακρίβεια: περισσότερα ή καθαρότερα δεδομένα μειώνουν την αβεβαιότητα του posterior.

## 7.4 Σύνδεση MAP: BLR ως Κανονικοποιημένα Ελάχιστα Τετράγωνα

Ο MAP εκτιμητής είναι το mode του posterior. Εφόσον το posterior είναι Γκαουσιανό,

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \boldsymbol{\mu}_N.$$

Ισοδύναμα,

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_w^2}\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\|_2^2 + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_0)^\mathsf{T}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_0).$$

Για $\boldsymbol{\mu}_0 = \mathbf{0}$ και $\boldsymbol{\Sigma}_0 = \alpha^{-1}\boldsymbol{I}$, προκύπτει η ridge παλινδρόμηση:

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2, \qquad \lambda = \sigma_w^2\alpha.$$

Έτσι καθίσταται ρητή η αντιστοιχία "πρότερη κατανομή $\leftrightarrow$ κανονικοποιητής".

## 7.5 Προβλεπτικές Κατανομές

### Πρότερη πρόβλεψη (πριν την παρατήρηση δεδομένων)

Για ένα νέο σημείο εισόδου $\boldsymbol{x}_*$ με διάνυσμα χαρακτηριστικών $\boldsymbol{\phi}_* = \boldsymbol{\phi}(\boldsymbol{x}_*)$, το υπό συνθήκη μοντέλο είναι

$$p(y_* \mid \boldsymbol{x}_*, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\phi}_*^\mathsf{T}\boldsymbol{\theta}, \sigma_w^2).$$

Πριν την παρατήρηση δεδομένων, η προβλεπτική κατανομή προκύπτει με μέση τιμή ως προς την πρότερη κατανομή:

$$p(y_* \mid \boldsymbol{x}_*) = \int p(y_* \mid \boldsymbol{x}_*, \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

Θέτοντας $z_* = \boldsymbol{\phi}_*^\mathsf{T}\boldsymbol{\theta}$, και δεδομένου ότι $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ και το $z_*$ είναι γραμμικός συνδυασμός,

$$z_* \sim \mathcal{N}\left(\boldsymbol{\phi}_*^\mathsf{T}\boldsymbol{\mu}_0,\ \boldsymbol{\phi}_*^\mathsf{T}\boldsymbol{\Sigma}_0\boldsymbol{\phi}_*\right).$$

Με ανεξάρτητο προσθετικό θόρυβο $y_* = z_* + w_*$, $w_* \sim \mathcal{N}(0, \sigma_w^2)$, παίρνουμε

$$\boxed{p(y_* \mid \boldsymbol{x}_*) = \mathcal{N}\left(\boldsymbol{\phi}_*^\mathsf{T}\boldsymbol{\mu}_0,\ \boldsymbol{\phi}_*^\mathsf{T}\boldsymbol{\Sigma}_0\boldsymbol{\phi}_* + \sigma_w^2\right).}$$

**Εκ των υστέρων πρόβλεψη (μετά την παρατήρηση δεδομένων)**

Μετά την παρατήρηση του συνόλου δεδομένων $\mathcal{D}$, λαμβάνουμε μέσο ως προς την εκ των υστέρων κατανομή:

$$p(y_* \mid \boldsymbol{x}_*, \boldsymbol{y}, \boldsymbol{\Phi}) = \int p(y_* \mid \boldsymbol{x}_*, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{\Phi})\, d\boldsymbol{\theta}.$$

Χρησιμοποιώντας $p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{\Phi}) = \mathcal{N}(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$, το ίδιο γραμμικό–Γκαουσιανό επιχείρημα οδηγεί σε

$$\boxed{p(y_* \mid \boldsymbol{x}_*, \boldsymbol{y}, \boldsymbol{\Phi}) = \mathcal{N}\big(\boldsymbol{\phi}_*^\mathsf{T}\boldsymbol{\mu}_N,\ \boldsymbol{\phi}_*^\mathsf{T}\boldsymbol{\Sigma}_N\boldsymbol{\phi}_* + \sigma_w^2\big)\,.}$$

**Αποσύνθεση μέσου και διακύμανσης.** Ο προβλεπτικός μέσος είναι

$$\mathbb{E}[y_* \mid \cdot] = \boldsymbol{\phi}_*^\mathsf{T}\boldsymbol{\mu}_N,$$

ενώ η προβλεπτική διακύμανση διασπάται σε δύο όρους:

$$\mathrm{Var}(y_* \mid \cdot) = \underbrace{\boldsymbol{\phi}_*^\mathsf{T}\boldsymbol{\Sigma}_N\boldsymbol{\phi}_*}_{\text{αβεβαιότητα παραμέτρων}} + \underbrace{\sigma_w^2}_{\text{θόρυβος παρατήρησης}}.$$

Καθώς λαμβάνονται περισσότερα ή πιο πληροφοριακά δεδομένα, η $\boldsymbol{\Sigma}_N$ συρρικνώνεται και τα διαστήματα αβεβαιότητας στενεύουν: ο μη αναγώγιμος όρος $\sigma_w^2$ παραμένει ως θόρυβος μέτρησης.

## 7.6 Πρακτική Σημείωση: Δειγματοληψία από την Πρότερη Κατανομή (Διαίσθηση στον Χώρο Συναρτήσεων)

Πριν την παρατήρηση δεδομένων, η Μπεϋζιανή γραμμική παλινδρόμηση ορίζει μια κατανομή πάνω σε συναρτήσεις μέσω δειγματοληψίας των βαρών:

$$\boldsymbol{\theta}^{(s)} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \qquad z^{(s)}(\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\theta}^{(s)}.$$

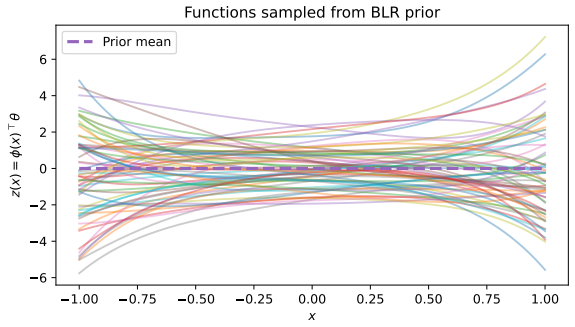
Προσθέτοντας θόρυβο παρατήρησης, προκύπτουν συνθετικά σύνολα δεδομένων:

$$y_i^{(s)} = z^{(s)}(\boldsymbol{x}_i) + w_i^{(s)}, \qquad w_i^{(s)} \sim \mathcal{N}(0, \sigma_w^2).$$

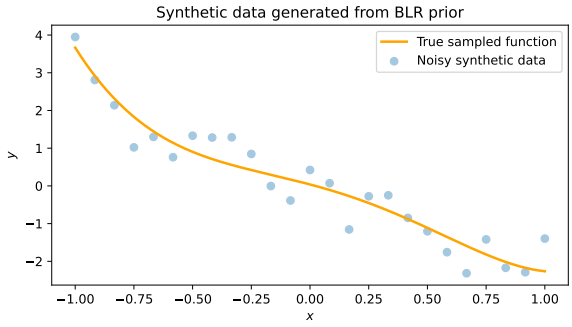
Αυτός είναι ένας χρήσιμος τρόπος οπτικοποίησης και ελέγχου των πρότερων κατανομών και των χαρακτηριστικών: δείχνει ποιες συναρτήσεις το μοντέλο θεωρεί εύλογες *πριν* παρατηρηθούν δεδομένα.

Το Σχ. 7.1 απεικονίζει μια σημαντική διαίσθηση πίσω από τη Μπεϋζιανή γραμμική παλινδρόμηση: πριν παρατηρηθούν δεδομένα, το μοντέλο ορίζει ήδη μια *κατανομή πάνω σε συναρτήσεις*. Η κατανομή αυτή επάγεται από την πρότερη κατανομή $p(\boldsymbol{\theta})$ σε συνδυασμό με την επιλεγμένη απεικόνιση χαρακτηριστικών $\boldsymbol{\phi}(\cdot)$.

Αριστερά βλέπουμε πολλαπλά δείγματα συναρτήσεων που προκύπτουν αν πρώτα δειγματοληφθεί $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ και στη συνέχεια υπολογιστεί η αντίστοιχη πρόβλεψη χωρίς θόρυβο $z(\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x})^\mathsf{T}\boldsymbol{\theta}$ σε όλο το πεδίο εισόδων. Οι καμπύλες αυτές αναπαριστούν το σύνολο των συναρτήσεων που η πρότερη κατανομή θεωρεί εύλογες *πριν* την παρατήρηση οποιωνδήποτε δεδομένων. Η επιλογή των χαρακτηριστικών και η πρότερη συνδιακύμανση $\boldsymbol{\Sigma}_0$ επηρεάζουν έντονα την ομαλότητα, το πλάτος και τη μεταβλητότητά τους.




Δείγματα συναρτήσεων από την πρότερη κατανομή




Ένα συνθετικό σύνολο δεδομένων

Σχήμα 7.1: Διαίσθηση της Μπεϋζιανής γραμμικής παλινδρόμησης μέσω δειγματοληψίας από την πρότερη κατανομή. Αριστερά: δειγματοληψία βαρών $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ και απεικόνιση των αντίστοιχων συναρτήσεων χωρίς θόρυβο $z(\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\theta}$, ώστε να οπτικοποιηθεί ποιες συναρτήσεις θεωρεί εύλογες η πρότερη κατανομή. Δεξιά: δειγματοληψία μίας τέτοιας συνάρτησης και προσθήκη Γκαουσιανού θορύβου $w \sim \mathcal{N}(0, \sigma_w^2)$ για τη δημιουργία ενός συνθετικού συνόλου εκπαίδευσης.

Δεξιά βλέπουμε πώς μπορούν να παραχθούν συνθετικά σύνολα δεδομένων από αυτή την πρότερη κατανομή. Μία συνάρτηση δειγματοληπτείται και προστίθεται Γκαουσιανός θόρυβος παρατήρησης $w \sim \mathcal{N}(0, \sigma_w^2)$ ώστε να παραχθούν σημεία εκπαίδευσης. Η διαδικασία αυτή αντιστοιχεί ακριβώς στο γενετικό (generative) μοντέλο που υποθέτει η Μπεϋζιανή γραμμική παλινδρόμηση.

Η δειγματοληψία από την πρότερη κατανομή είναι χρήσιμη για πολλούς λόγους. Αποτελεί έναν έλεγχο λογικής για τον σχεδιασμό των χαρακτηριστικών και την επιλογή της πρότερης κατανομής, επιτρέποντάς μας να επιβεβαιώσουμε ότι το μοντέλο ενσωματώνει εύλογες υποθέσεις πριν την προσαρμογή στα δεδομένα. Επιπλέον, βοηθά στην κατανόηση της αβεβαιότητας: ακόμη και χωρίς δεδομένα, οι προβλέψεις διαφέρουν επειδή πολλές τιμές παραμέτρων είναι πιθανές. Τέλος, αποτελεί έναν πρακτικό τρόπο δημιουργίας ελεγχόμενων συνθετικών συνόλων δεδομένων για πειράματα, επιδείξεις ή ασκήσεις, υπό τις ίδιες ακριβώς υποθέσεις που χρησιμοποιούνται στη διαδικασία εκτίμησης.

## 7.7 Παράδειγμα: Μπεϋζιανή Γραμμική Παλινδρόμηση (1Δ, Κλειστή Μορφή)

Θα εξετάσουμε ένα πλήρες παράδειγμα Μπεϋζιανής γραμμικής παλινδρόμησης για ένα απλό μονοδιάστατο γραμμικό μοντέλο (σταθερός όρος + κλίση). Στόχος είναι ο υπολογισμός της εκ των υστέρων κατανομής των βαρών και της εκ των υστέρων προβλεπτικής κατανομής σε ένα νέο σημείο εισόδου.

**Μοντέλο και χαρακτηριστικά.** Έστω

$$\boldsymbol{\phi}(x) = \begin{bmatrix} 1 \\ x \end{bmatrix}, \qquad y_i = \boldsymbol{\phi}(x_i)^\top \boldsymbol{\theta} + w_i, \qquad w_i \sim \mathcal{N}(0, \sigma_w^2), \qquad \boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}.$$

Υποθέτουμε διακύμανση θορύβου

$$\sigma_w^2 = 1.$$

**Πρότερη κατανομή.** Χρησιμοποιούμε μια ισότροπη Γκαουσιανή πρότερη κατανομή μηδενικού μέσου:

$$p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \alpha^{-1}\boldsymbol{I}), \qquad \alpha = 1 \quad \Rightarrow \quad \boldsymbol{\Sigma}_0 = \boldsymbol{I},\ \boldsymbol{\mu}_0 = \mathbf{0}.$$

**Δεδομένα εκπαίδευσης.** Χρησιμοποιούμε τρεις παρατηρήσεις:

$$(x_1, y_1) = (0, 1), \quad (x_2, y_2) = (1, 2), \quad (x_3, y_3) = (2, 2).$$

Τότε

$$\boldsymbol{\Phi} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}.$$

**Βήμα 1: Υπολογισμός της εκ των υστέρων συνδιακύμανσης $\boldsymbol{\Sigma}_N$**

Η εκ των υστέρων συνδιακύμανση της BLR είναι

$$\boldsymbol{\Sigma}_N = \left( \boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma_w^2} \boldsymbol{\Phi}^\mathsf{T} \boldsymbol{\Phi} \right)^{-1}.$$

Υπολογίζουμε

$$\boldsymbol{\Sigma}_0^{-1} = \boldsymbol{I}, \qquad \boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi} = \begin{bmatrix} 3 & 3 \\ 3 & 5 \end{bmatrix}.$$

Εφόσον $\sigma_w^2 = 1$,

$$\boldsymbol{\Sigma}_N^{-1} = \boldsymbol{I} + \boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Phi} = \begin{bmatrix} 4 & 3 \\ 3 & 6 \end{bmatrix}.$$

Αντιστρέφοντας:

$$\det\left(\boldsymbol{\Sigma}_N^{-1}\right) = 4 \cdot 6 - 3 \cdot 3 = 24 - 9 = 15,$$

$$\Rightarrow \quad \boldsymbol{\Sigma}_N = (\boldsymbol{\Sigma}_N^{-1})^{-1} = \frac{1}{15} \begin{bmatrix} 6 & -3 \\ -3 & 4 \end{bmatrix}.$$

**Βήμα 2: Υπολογισμός του εκ των υστέρων μέσου $\boldsymbol{\mu}_N$**

Ο εκ των υστέρων μέσος είναι

$$\boldsymbol{\mu}_N = \boldsymbol{\Sigma}_N \left( \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \frac{1}{\sigma_w^2} \boldsymbol{\Phi}^\mathsf{T} \boldsymbol{y} \right).$$

Εδώ $\boldsymbol{\mu}_0 = \mathbf{0}$ και $\sigma_w^2 = 1$, άρα

$$\boldsymbol{\mu}_N = \boldsymbol{\Sigma}_N \boldsymbol{\Phi}^\mathsf{T} \boldsymbol{y}.$$

Υπολογίζουμε

$$\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{y} = \begin{bmatrix} 1 + 2 + 2 \\ 0 \cdot 1 + 1 \cdot 2 + 2 \cdot 2 \end{bmatrix} = \begin{bmatrix} 5 \\ 6 \end{bmatrix}.$$

Άρα

$$\boldsymbol{\mu}_N = \frac{1}{15} \begin{bmatrix} 6 & -3 \\ -3 & 4 \end{bmatrix} \begin{bmatrix} 5 \\ 6 \end{bmatrix} = \frac{1}{15} \begin{bmatrix} 30 - 18 \\ -15 + 24 \end{bmatrix} = \begin{bmatrix} 4/5 \\ 3/5 \end{bmatrix}.$$

Επομένως,

$$\boxed{p(\boldsymbol{\theta} \mid \mathcal{D}) = \mathcal{N}\left( \begin{bmatrix} 4/5 \\ 3/5 \end{bmatrix}, \frac{1}{15} \begin{bmatrix} 6 & -3 \\ -3 & 4 \end{bmatrix} \right).}$$

**Βήμα 3: Εκ των υστέρων προβλεπτική κατανομή σε νέο σημείο**

Θεωρούμε ένα σημείο ελέγχου $x_* = 1.5$, οπότε

$$\boldsymbol{\phi}_* = \boldsymbol{\phi}(1.5) = \begin{bmatrix} 1 \\ 1.5 \end{bmatrix}.$$

Η εκ των υστέρων προβλεπτική κατανομή είναι Γκαουσιανή:

$$p(y_* \mid x_*, \mathcal{D}) = \mathcal{N}\big(m_*,\ s_*^2\big)\,,$$

με

$$m_* = \boldsymbol{\phi}_*^{\mathsf{T}} \boldsymbol{\mu}_N, \qquad s_*^2 = \boldsymbol{\phi}_*^{\mathsf{T}} \boldsymbol{\Sigma}_N \boldsymbol{\phi}_* + \sigma_w^2.$$

Υπολογισμός προβλεπτικού μέσου:

$$m_* = \begin{bmatrix} 1 & 1.5 \end{bmatrix} \begin{bmatrix} 4/5 \\ 3/5 \end{bmatrix} = \frac{4}{5} + \frac{4.5}{5} = \frac{8.5}{5} = 1.7.$$

Υπολογισμός όρου αβεβαιότητας παραμέτρων:

$$\boldsymbol{\Sigma}_N \boldsymbol{\phi}_* = \frac{1}{15} \begin{bmatrix} 6 & -3 \\ -3 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 1.5 \end{bmatrix} = \frac{1}{15} \begin{bmatrix} 6 - 4.5 \\ -3 + 6 \end{bmatrix} = \frac{1}{15} \begin{bmatrix} 1.5 \\ 3 \end{bmatrix} = \begin{bmatrix} 0.1 \\ 0.2 \end{bmatrix}.$$

Άρα

$$\boldsymbol{\phi}_*^{\mathsf{T}} \boldsymbol{\Sigma}_N \boldsymbol{\phi}_* = \begin{bmatrix} 1 & 1.5 \end{bmatrix} \begin{bmatrix} 0.1 \\ 0.2 \end{bmatrix} = 0.1 + 0.3 = 0.4.$$

Τελικά,

$$s_*^2 = 0.4 + \sigma_w^2 = 0.4 + 1 = 1.4.$$

Άρα

$$\boxed{p(y_* \mid x_* = 1.5, \mathcal{D}) = \mathcal{N}(1.7,\ 1.4).}$$

**Ερμηνεία.**

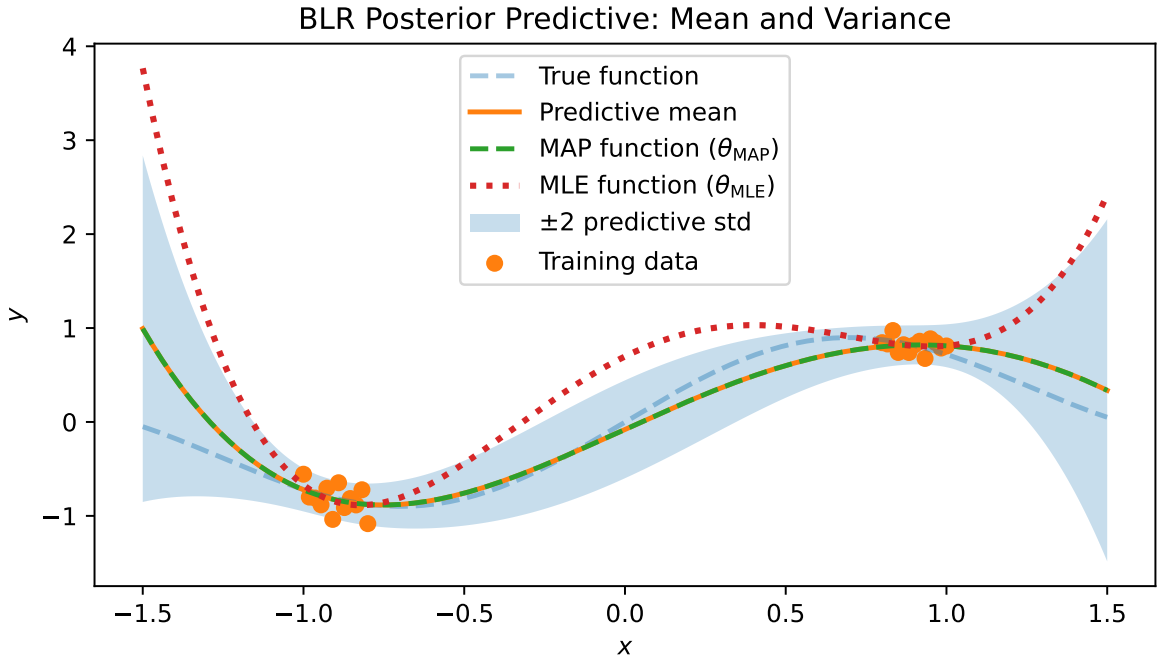
- Ο εκ των υστέρων μέσος $\boldsymbol{\mu}_N$ είναι ο MAP εκτιμητής.
- Η εκ των υστέρων συνδιακύμανση $\boldsymbol{\Sigma}_N$ ποσοτικοποιεί την αβεβαιότητα στα βάρη και συνήθως μειώνεται καθώς προστίθενται δεδομένα.
- Η προβλεπτική διακύμανση διασπάται σε $\boldsymbol{\phi}_*^{\mathsf{T}} \boldsymbol{\Sigma}_N \boldsymbol{\phi}_*$ (αβεβαιότητα παραμέτρων) και $\sigma_w^2$ (μη αναγώγιμος θόρυβος μέτρησης).

## 7.8 Υλοποίηση σε Python

Ο κώδικας κατασκευάζει ένα συνθετικό πρόβλημα παλινδρόμησης όπου η υποκείμενη συνάρτηση είναι μη γραμμική, αλλά σκόπιμα προσαρμόζουμε ένα *μοντέλο πολυωνυμικών χαρακτηριστικών* βαθμού $P$:

$$\phi(x) = \begin{bmatrix} 1 & x & x^2 & \cdots & x^P \end{bmatrix}^{\mathsf{T}}, \qquad y \approx \boldsymbol{\theta}^{\mathsf{T}} \phi(x) + w, \qquad w \sim \mathcal{N}(0, \sigma_w^2).$$

Αυτό είναι ένα τυπικό παράδειγμα «γραμμικού μοντέλου πάνω σε μη γραμμικά χαρακτηριστικά»: το μοντέλο είναι γραμμικό ως προς $\boldsymbol{\theta}$, αλλά μη γραμμικό ως προς το $x$.




Σχήμα 7.2: Εκ των υστέρων προβλεπτική συμπεριφορά στη Μπεϋζιανή γραμμική παλινδρόμηση (BLR), σε σύγκριση με τις σημειακές εκτιμήσεις MLE και MAP. Τα δεδομένα εκπαίδευσης είναι συγκεντρωμένα σε δύο απομακρυσμένες περιοχές εισόδου (κοντά στο $x \approx -1$ και στο $x \approx +1$), αφήνοντας ένα κενό χωρίς παρατηρήσεις. Το σχήμα δείχνει: (i) τον *προβλεπτικό μέσο* της BLR (εκ των υστέρων προβλεπτικός μέσος), (ii) μια ζώνη αβεβαιότητας $\pm 2$ προγνωστικών τυπικών αποκλίσεων, (iii) τη *MAP συνάρτηση* που επάγεται από τα βάρη $\boldsymbol{\theta}_{\text{MAP}}$ (για Γκαουσιανό posterior συμπίπτει με τον posterior μέσο), και (iv) τη *MLE συνάρτηση* που επάγεται από $\boldsymbol{\theta}_{\text{MLE}}$ (χωρίς πρότερη κατανομή).

**Σχεδιασμός συνόλου δεδομένων: γιατί το κενό έχει σημασία.** Οι είσοδοι δειγματοληπτούνται από δύο ασύνδετα διαστήματα:

$$x \in [-1, -0.8] \ \cup \ [0.8, 1],$$

οπότε το μοντέλο δεν βλέπει καθόλου δεδομένα στη μεσαία περιοχή. Ακριβώς εκεί η Μπεϋζιανή προβλεπτική αβεβαιότητα γίνεται ιδιαίτερα πληροφοριακή: η BLR πρέπει να είναι βέβαιη όπου υπάρχουν δεδομένα και λιγότερο βέβαιη όπου δεν υπάρχουν.

**Τρεις διαφορετικοί προβλεπτές στο γράφημα.**

- **Συνάρτηση MLE** (σημειακή εκτίμηση, χωρίς πρότερη κατανομή). Ο κώδικας υπολογίζει
$$\boldsymbol{\theta}_{\text{MLE}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{y} \mid \boldsymbol{\theta}) \quad \Longleftrightarrow \quad \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\|_2^2,$$
υλοποιημένο ως $\boldsymbol{\theta}_{\text{MLE}} = \text{pinv}(\boldsymbol{\Phi})\,\boldsymbol{y}$. Αυτό αποδίδει μία μοναδική καμπύλη $\hat{y}(x) = \phi(x)^{\mathsf{T}}\boldsymbol{\theta}_{\text{MLE}}$ *χωρίς καμία ποσοτικοποίηση αβεβαιότητας*.

- **Συνάρτηση MAP** (σημειακή εκτίμηση με πρότερη κατανομή). Η πρότερη κατανομή είναι Γκαουσιανή,
$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \qquad \boldsymbol{\Sigma}_0 = \alpha^{-1}\boldsymbol{I},$$

γεγονός που οδηγεί σε MAP εκτιμητή ισοδύναμο με ridge παλινδρόμηση:

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_w^2}\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\|_2^2 + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_0)^{\mathsf{T}}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_0).$$

Ο κώδικας υπολογίζει το $\boldsymbol{\theta}_{\mathrm{MAP}}$ ρητά σε κλειστή μορφή. Και εδώ προκύπτει μία μοναδική καμπύλη, συνήθως πιο ομαλή και λιγότερο ακραία από τη MLE λόγω της συρρίκνωσης (shrinkage).

- **Εκ των υστέρων προβλεπτική BLR** (κατανομή πάνω στις εξόδους). Η BLR δεν επιστρέφει μία μόνο συνάρτηση, αλλά μια προβλεπτική κατανομή:

$$p(y_* \mid x_*, \mathcal{D}) = \mathcal{N}\Big( \underbrace{\phi(x_*)^{\mathsf{T}}\boldsymbol{\mu}_N}_{\text{προβλεπτικός μέσος}} \;,\; \underbrace{\phi(x_*)^{\mathsf{T}}\boldsymbol{\Sigma}_N\phi(x_*)}_{\text{αβεβαιότητα παραμέτρων}} + \underbrace{\sigma_w^2}_{\text{θόρυβος}} \Big).$$

Η σκιασμένη ζώνη στο γράφημα αντιστοιχεί σε $\pm 2$ προβλεπτικές τυπικές αποκλίσεις, δηλαδή περίπου σε διάστημα εμπιστοσύνης 95% υπό το Γκαουσιανό προβλεπτικό μοντέλο.

**Κύρια συμπεράσματα.**

- **Οι σημειακοί εκτιμητές (MLE/MAP) δίνουν καμπύλες αλλά όχι αβεβαιότητα.** Δεν μπορούν να εκφράσουν ότι η πρόβλεψη στο κενό είναι λιγότερο αξιόπιστη από ό,τι κοντά στις περιοχές όπου υπάρχουν δεδομένα.
- **Η αβεβαιότητα της BLR αυξάνεται μακριά από τα δεδομένα.** Σε περιοχές με λίγα ή καθόλου σημεία εκπαίδευσης, ο όρος $\phi(x)^{\mathsf{T}}\boldsymbol{\Sigma}_N\phi(x)$ αυξάνεται, διευρύνοντας τη ζώνη αβεβαιότητας.
- **Η διαφορά MAP–MLE είναι εντονότερη όταν τα δεδομένα είναι λίγα ή τα χαρακτηριστικά πολύ εύκαμπτα.** Η πρότερη κατανομή συρρικνώνει τους συντελεστές και αποτρέπει ακραία πολυωνυμική συμπεριφορά, οπότε η καμπύλη MAP είναι συχνά λιγότερο ταλαντωτική από τη MLE (ιδίως εκτός του εύρους παρατηρήσεων).

**Λεπτομέρειες υλοποίησης που αντικατοπτρίζονται στον κώδικα.** Ο κώδικας χρησιμοποιεί τον posterior σε κλειστή μορφή:

$$\boldsymbol{\Sigma}_N^{-1} = \boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma_w^2}\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi}, \qquad \boldsymbol{\mu}_N = \boldsymbol{\Sigma}_N\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \frac{1}{\sigma_w^2}\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{y}\right),$$

και στη συνέχεια υπολογίζει τον προβλεπτικό μέσο και τη διακύμανση σε ένα πλέγμα σημείων μέσω

$$\mathrm{Var}(y_* \mid \cdot) = \mathrm{diag}(\boldsymbol{\Phi}_*\boldsymbol{\Sigma}_N\boldsymbol{\Phi}_*^{\mathsf{T}}) + \sigma_w^2,$$

υλοποιημένο με το αποδοτικό τέχνασμα της διαγωνίου:

$$\mathrm{diag}(\boldsymbol{\Phi}_*\boldsymbol{\Sigma}_N\boldsymbol{\Phi}_*^{\mathsf{T}}) = \sum_j (\boldsymbol{\Phi}_*\boldsymbol{\Sigma}_N)_{\cdot j} \odot (\boldsymbol{\Phi}_*)_{\cdot j}.$$

Αυτός είναι ο λόγος που ο κώδικας υπολογίζει `param_var = np.sum(Phi_grid @ SigmaN * Phi_grid, axis=1)`.

```python
import numpy as np
import matplotlib.pyplot as plt

# ---------------------------
# 1) Synthetic dataset
# ---------------------------
np.random.seed(1)

# N = 25
# x_data = np.linspace(-1., 1., N)
N = 25
N_left = N // 2
N_right = N - N_left

x_left  = np.linspace(-1.0, -0.8, N_left, endpoint=True)
x_right = np.linspace(0.8,  1.0, N_right, endpoint=True)

x_data = np.concatenate([x_left, x_right])
sigma_w = 0.1    # noise std

# True underlying function (just for data gen)
def f_true(x):
    return 0.7*np.sin(2.5*x) + 0.3*x

y_data = f_true(x_data) + sigma_w*np.random.randn(N)

# ---------------------------
# 2) Feature map (polynomial)
# ---------------------------
P = 4  # order -> feature dim = P+1

def phi(x, P):
    return np.vstack([x**k for k in range(P+1)]).T

Phi_data = phi(x_data, P)  # (N, P+1)

# Grid for prediction
x_grid = np.linspace(-1.5, 1.5, 300)
Phi_grid = phi(x_grid, P)

y_grid = f_true(x_grid)

# ---------------------------
# 3) Prior over theta
# ---------------------------
alpha = 2.0
mu0 = np.zeros(P+1)
Sigma0 = (1/alpha) * np.eye(P+1)

# ---------------------------
# 4) Posterior (closed form)
# ---------------------------
SigmaN_inv = np.linalg.inv(Sigma0) + (1/sigma_w**2) * (Phi_data.T @
    Phi_data)
SigmaN = np.linalg.inv(SigmaN_inv)

muN = SigmaN @ (np.linalg.inv(Sigma0) @ mu0 + (1/sigma_w**2) * Phi_data
```

```
    .T @ y_data)

# ---------------------------
# 4b) MAP estimate for theta
# ---------------------------
# For Gaussian posterior, MAP = posterior mean
# theta_map = muN.copy()

# (Optional explicit ridge/MAP form; should match muN)
theta_map = np.linalg.inv(Phi_data.T @ Phi_data + sigma_w**2 * np.
    linalg.inv(Sigma0)) @ (Phi_data.T @ y_data + sigma_w**2 * np.linalg.
    inv(Sigma0) @ mu0)

# ---------------------------
# 4c) MLE estimate for theta (no prior)
# ---------------------------
# With N < P+1, use pseudoinverse for stability
theta_mle = np.linalg.pinv(Phi_data) @ y_data
y_mle = Phi_grid @ theta_mle

# Function induced by MAP parameters
y_map = Phi_grid @ theta_map

# ---------------------------
# 5) Posterior predictive
# ---------------------------
# Mean
y_mean = Phi_grid @ muN

# Variance = parameter uncertainty + noise
param_var = np.sum(Phi_grid @ SigmaN * Phi_grid, axis=1)  # diag(Phi Σ
    Phi^T)
pred_var = param_var + sigma_w**2
pred_std = np.sqrt(pred_var)

# ---------------------------
# 6) Plot
# ---------------------------
plt.figure(figsize=(6.8, 4.0))

# the actual function
plt.plot(x_grid, y_grid, "--", linewidth=2, alpha=0.4, label="True
    function")

# predictive mean
plt.plot(x_grid, y_mean, linewidth=2, label="Predictive mean")

# MAP function
plt.plot(x_grid, y_map, "--", linewidth=2, label=r"MAP function ($\
    theta_{\mathrm{MAP}}$)")

# MLE function
plt.plot(x_grid, y_mle, ":", linewidth=2.5, label=r"MLE function ($\
    theta_{\mathrm{MLE}}$)")

# uncertainty band ±2 std
plt.fill_between(
    x_grid,
```

```
    y_mean - 2*pred_std,
    y_mean + 2*pred_std,
    alpha=0.25,
    label=r"$\pm 2$ predictive std"
)

# data
plt.scatter(x_data, y_data, s=35, label="Training data")

plt.xlabel(r"$x$")
plt.ylabel(r"$y$")
plt.title("BLR Posterior Predictive: Mean and Variance")
plt.legend()
plt.tight_layout()
plt.show()

print("theta_MAP =", theta_map)
print("theta_MLE =", theta_mle)
```

Listing 2: Bayesian Linear Regression

## 7.9 Διαδοχική Μπεϋζιανή Γραμμική Παλινδρόμηση (Αναδρομική Ενημέρωση)

Ένα σημαντικό πρακτικό πλεονέκτημα της Μπεϋζιανής γραμμικής παλινδρόμησης είναι ότι μπορεί να ενημερώνεται *online*: όταν καταφθάνει ένα νέο δείγμα, ενημερώνουμε το posterior χωρίς να ξανα-επεξεργαστούμε ολόκληρο το σύνολο δεδομένων. Έστω ότι μετά την παρατήρηση

$$\mathcal{D}_k = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{k},$$

το posterior μας είναι

$$p(\boldsymbol{\theta} \mid \mathcal{D}_k) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Ένα νέο σημείο δεδομένων $(\boldsymbol{x}_{k+1}, y_{k+1})$ επάγει ένα διάνυσμα χαρακτηριστικών

$$\boldsymbol{\phi}_{k+1} \equiv \boldsymbol{\phi}(\boldsymbol{x}_{k+1}), \qquad y_{k+1} = \boldsymbol{\phi}_{k+1}^{\mathsf{T}}\boldsymbol{\theta} + w_{k+1}, \qquad w_{k+1} \sim \mathcal{N}(0, \sigma_w^2).$$

Ο κανόνας του Bayes λέει ότι το ενημερωμένο posterior είναι ανάλογο του προηγούμενου posterior επί τη νέα πιθανοφάνεια:

$$p(\boldsymbol{\theta} \mid \mathcal{D}_{k+1}) \propto p(y_{k+1} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D}_k).$$

Επειδή και οι δύο παράγοντες είναι Γκαουσιανοί ως προς $\boldsymbol{\theta}$, το posterior παραμένει Γκαουσιανό:

$$p(\boldsymbol{\theta} \mid \mathcal{D}_{k+1}) = \mathcal{N}(\boldsymbol{\mu}_{k+1}, \boldsymbol{\Sigma}_{k+1}).$$

Είναι ιδιαίτερα κομψό να εκφράσουμε την αναδρομή στη *μορφή πληροφορίας* (information form). Ορίζουμε τον πίνακα ακρίβειας (πληροφορίας)

$$\boldsymbol{\Lambda}_k \triangleq \boldsymbol{\Sigma}_k^{-1}.$$

Τότε κάθε νέο δείγμα συνεισφέρει μια αύξηση πληροφορίας τάξης 1:

$$\boxed{\boldsymbol{\Lambda}_{k+1} = \boldsymbol{\Lambda}_k + \frac{1}{\sigma_w^2}\boldsymbol{\phi}_{k+1}\boldsymbol{\phi}_{k+1}^{\mathsf{T}}.}$$

Αντίστοιχα, ορίζουμε το διάνυσμα πληροφορίας

$$\boldsymbol{\eta}_k \triangleq \boldsymbol{\Lambda}_k \boldsymbol{\mu}_k.$$

Η ενημέρωση του μέσου γίνεται ένας απλός αθροιστικός κανόνας στον χώρο πληροφορίας:

$$\boxed{\boldsymbol{\eta}_{k+1} = \boldsymbol{\eta}_k + \frac{1}{\sigma_w^2}\boldsymbol{\phi}_{k+1} y_{k+1}, \qquad \boldsymbol{\mu}_{k+1} = \boldsymbol{\Lambda}_{k+1}^{-1}\boldsymbol{\eta}_{k+1}.}$$

Αυτή η αναπαράσταση αναδεικνύει την ερμηνεία της διαδοχικής BLR: συσσωρεύουμε διαρκώς πληροφορία για το $\boldsymbol{\theta}$. Όταν η διακύμανση θορύβου παρατήρησης $\sigma_w^2$ είναι μικρή, κάθε δείγμα θεωρείται πιο αξιόπιστο και προσθέτει περισσότερη πληροφορία· όταν $\sigma_w^2$ είναι μεγάλη, η ενημέρωση είναι μικρότερη.

Η ίδια αναδρομή μπορεί επίσης να γραφτεί σε μορφή τύπου φίλτρου Kalman, η οποία ίσως είναι πιο διαισθητική αν έχετε δει αναδρομικά ελάχιστα τετράγωνα ή ενημερώσεις Kalman. Η προβλεπτική κατανομή της νέας εξόδου είναι

$$y_{k+1} \mid \mathcal{D}_k \sim \mathcal{N}\big(\boldsymbol{\phi}_{k+1}^{\mathsf{T}}\boldsymbol{\mu}_k,\ s_{k+1}\big), \qquad s_{k+1} \triangleq \boldsymbol{\phi}_{k+1}^{\mathsf{T}}\boldsymbol{\Sigma}_k\boldsymbol{\phi}_{k+1} + \sigma_w^2.$$

Ορίζουμε την καινοτομία (σφάλμα πρόβλεψης)

$$e_{k+1} \triangleq y_{k+1} - \boldsymbol{\phi}_{k+1}^{\mathsf{T}}\boldsymbol{\mu}_k,$$

και το διάνυσμα κέρδους

$$\boldsymbol{K}_{k+1} \triangleq \frac{\boldsymbol{\Sigma}_k\boldsymbol{\phi}_{k+1}}{s_{k+1}}.$$

Τότε ο εκ των υστέρων μέσος και η συνδιακύμανση ενημερώνονται ως

$$\boxed{\boldsymbol{\mu}_{k+1} = \boldsymbol{\mu}_k + \boldsymbol{K}_{k+1}e_{k+1}, \qquad \boldsymbol{\Sigma}_{k+1} = \boldsymbol{\Sigma}_k - \boldsymbol{K}_{k+1}\boldsymbol{\phi}_{k+1}^{\mathsf{T}}\boldsymbol{\Sigma}_k.}$$

Αυτό είναι ακριβώς η ενημέρωση μέτρησης ενός φίλτρου Kalman για μια *στατική* κατάσταση $\boldsymbol{\theta}$ (χωρίς δυναμική), με πίνακα μέτρησης $\boldsymbol{\phi}_{k+1}^{\mathsf{T}}$.

## 7.10 Τι προσθέτει η Μπεϋζιανή Γραμμική Παλινδρόμηση πέρα από LS / MLE

Τα ελάχιστα τετράγωνα (LS) και η μέγιστη πιθανοφάνεια (MLE) παρέχουν ένα *μοναδικό* διάνυσμα παραμέτρων βέλτιστης προσαρμογής. Αντίθετα, η Μπεϋζιανή γραμμική παλινδρόμηση (BLR) παράγει μια πλήρη εκ των υστέρων κατανομή

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \mathcal{N}(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N),$$

η οποία ποσοτικοποιεί άμεσα την αβεβαιότητά μας για κάθε παράμετρο, καθώς και τις συσχετίσεις μεταξύ των παραμέτρων.

Ένα δεύτερο πλεονέκτημα είναι ότι η BLR ενσωματώνει την κανονικοποίηση με θεμελιωμένο τρόπο μέσω της πρότερης κατανομής. Για Γκαουσιανό πρότερη κατανομή, ο MAP εκτιμητής συμπίπτει με τον εκ των υστέρων μέσο και είναι ισοδύναμος με ridge παλινδρόμηση:

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2\sigma_w^2}\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\|_2^2 + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_0)^{\mathsf{T}}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_0).$$

Έτσι, η πρότερη κατανομή ελέγχει την πολυπλοκότητα του μοντέλου και βελτιώνει τη αριθμητική σταθερότητα, ιδιαίτερα όταν τα δεδομένα είναι λίγα ή η διάσταση των χαρακτηριστικών είναι μεγάλη.

Το σημαντικότερο πλεονέκτημα είναι ότι η BLR παρέχει *προβλεπτική αβεβαιότητα*. Για μια νέα είσοδο $\boldsymbol{x}_*$ με χαρακτηριστικά $\boldsymbol{\phi}_*$, η εκ των υστέρων προβλεπτική κατανομή είναι

$$p(y_* \mid \boldsymbol{x}_*, \mathcal{D}) = \mathcal{N}\big(\boldsymbol{\phi}_*^\mathsf{T}\boldsymbol{\mu}_N,\ \boldsymbol{\phi}_*^\mathsf{T}\boldsymbol{\Sigma}_N\boldsymbol{\phi}_* + \sigma_w^2\big)\,.$$

Η προβλεπτική διακύμανση αποσυντίθεται σε δύο διαισθητικά μέρη:

$$\mathrm{Var}(y_* \mid \cdot) = \underbrace{\boldsymbol{\phi}_*^\mathsf{T}\boldsymbol{\Sigma}_N\boldsymbol{\phi}_*}_{\text{αβεβαιότητα στις } \boldsymbol{\theta}} + \underbrace{\sigma_w^2}_{\text{θόρυβος παρατήρησης}}\,.$$

Ο πρώτος όρος μικραίνει όταν τα δεδομένα περιορίζουν ισχυρά τις παραμέτρους (συγκέντρωση του posterior), ενώ ο δεύτερος όρος είναι μη αναγώγιμος θόρυβος μέτρησης.

Τέλος, η BLR συνδέεται φυσικά με την ασυμπτωτική συμπεριφορά της MLE: καθώς το $N$ αυξάνεται και τα δεδομένα είναι πληροφοριακά, η εκ των υστέρων συνδιακύμανση συρρικνώνεται και ο εκ των υστέρων μέσος προσεγγίζει τη MLE υπό τυπικές συνθήκες κανονικότητας. Με αυτή την έννοια, η BLR μπορεί να ιδωθεί ως LS/MLE εμπλουτισμένη με πρότερη γνώση και έναν συνεκτικό μηχανισμό ποσοτικοποίησης της αβεβαιότητας.

# Glossary (English – Greek)

**estimation** εκτίμηση (παραμέτρων)

**identification** εκτίμηση (παραμέτρων)

**parameter** παράμετρος

**model** μοντέλο

**loss function** συνάρτηση κόστους

**cost function / objective** συνάρτηση κόστους

**optimization** βελτιστοποίηση

**gradient** παράγωγος

**Hessian** Εσσιανός

**Jacobian** Ιακωβιανός

**line search** αναζήτηση γραμμής

**damping** απόσβεση

**trust region** περιοχή εμπιστοσύνης

**least squares (LS)** ελάχιστα τετράγωνα

**nonlinear least squares** μη γραμμικά ελάχιστα τετράγωνα

**regularization** κανονικοποίηση

**ridge (Tikhonov) regularization** κανονικοποίηση ridge (Tikhonov)

**prior** πρότερη κατανομή

**likelihood** πιθανοφάνεια

**posterior** εκ των υστέρων κατανομή

**evidence / marginal likelihood** περιθώρια πιθανοφάνεια

**Bayesian inference** Μπεϋζιανή συμπερασματολογία

**Bayes' rule** κανόνας του Bayes

**maximum likelihood estimation (MLE)** μέγιστη πιθανοφάνεια

**maximum a posteriori (MAP)** μέγιστο εκ των υστέρων

**minimum mean square error (MMSE)** ελάχιστο μέσο τετραγωνικό σφάλμα

**minimum mean absolute error (MMAE)** ελάχιστο μέσο απόλυτο σφάλμα

**point estimate** σημειακή εκτίμηση

**posterior mean** εκ των υστέρων μέση τιμή

**posterior mode** εκ των υστέρων επικρατούσα τιμή

**posterior median** εκ των υστέρων διάμεσος

**feature** χαρακτηριστικό

**feature map** απεικόνιση χαρακτηριστικών

**basis function** συνάρτηση βάσης

**design matrix** πίνακας σχεδίασης

**autoregressive (AR) model** αυτοπαλίνδρομο μοντέλο

**ARMA model** αυτοπαλίνδρομο–κινούμενου μέσου μοντέλο

**nonlinear autoregressive (NAR) model** μη γραμμικό αυτοπαλίνδρομο μοντέλο

**prediction error** σφάλμα πρόβλεψης

**predictive distribution** προβλεπτική κατανομή

**predictive mean** προβλεπτική μέση τιμή

**predictive variance** προβλεπτική διασπορά

**Bayesian linear regression (BLR)** Μπεϋζιανή γραμμική παλινδρόμηση

**posterior predictive** εκ των υστέρων προβλεπτική κατανομή

**prior predictive** πρότερη προβλεπτική κατανομή

**uncertainty** αβεβαιότητα

**parameter uncertainty** αβεβαιότητα παραμέτρων

**observation noise** θόρυβος παρατήρησης

**innovation** καινοτομία (σφάλμα πρόβλεψης)

**precision matrix** πίνακας ακρίβειας

**information form** μορφή πληροφορίας

**Kalman filter** φίλτρο Kalman

**recursive update** αναδρομική ενημέρωση