Gaussian sequence models

Paul Delatte
delatte@usc.edu
University of Southern California

Last updated: 08 November, 2022

1 Gaussian sequence model

The **Gaussian sequence model** or **Gaussian model in sequence space** is given for $n \in \mathbb{N}$ by

$$Y_i = \theta_i + \frac{\sigma}{\sqrt{n}} Z_i, \quad i \in \mathcal{I} \subseteq \mathbb{N},$$

where the θ_i are unknown constants, $\sigma > 0$ is a known constant, and $Z_i \overset{iid}{\sim} N(0,1)$. In other words,

$$Y_i \sim N\left(\theta_i, \frac{\sigma^2}{n}\right),$$

that is, we observe a sequence (finite or infinite) $(Y_i)_{i \in \mathcal{I}}$ of Gaussian random variables with unknown means θ_i but known variance σ^2/n. Since the variables Z_i are independent, the variables Y_i are also independent. Note that there is something natural (namely averaging, for connecting the model to nonparametric regression) but nothing special about the scaling $\sqrt{n}$: a more general formulation of the Gaussian sequence model is given by

$$Y_i = \theta_i + \sigma \zeta_n Z_i, \quad i \in \mathcal{I} \subseteq \mathbb{N},$$

for some known constants ζ_n (which yields back the previous model for $\zeta_n = 1/\sqrt{n}$).

If $\mathcal{I} = \{1, \dots, n\}$, then the Gaussian sequence model is known as the **normal means model** and is given for $n \in \mathbb{N}$ by

$$Y_i = \theta_i + \frac{\sigma}{\sqrt{n}} Z_i, \quad i \in \{1, \dots, n\}.$$

That is,

$$Y_i \sim N\left(\theta_i, \frac{\sigma^2}{n}\right), \quad i \in \{1, \dots, n\}.$$

Since the variables Y_i are independent, the model rewrites again as a multivariate Gaussian

$$Y \sim N\left(\theta, \frac{\sigma^2}{n} I_n\right),$$

where $Y = (Y_1, \dots, Y_n)$ and $\theta = (\theta_1, \dots, \theta_n)$. Even if the normal means model is finite-dimensional for n fixed, it is in essence nonparametric for the number $|\mathcal{I}| = n$ of unknowns θ_i grows as fast as the number n of data available. The generalization to other scalings in this case takes the form $Y \sim N(\theta, \zeta_i^2 \sigma^2 I_n)$. This includes for $\zeta_i = 1$ the simple multivariate Gaussian mean model given by $Y \sim N(\theta, \sigma^2 I_n)$ (in which Stein's phenomenon is usually exhibited).

2 Gaussian white noise model

The **Gaussian white noise model** is given for $n \in \mathbb{N}$ by the stochastic differential equation

$$dY(t) = f(t)\,dt + \frac{\sigma}{\sqrt{n}} dW(t), \quad t \in [0,1],$$

where $f \in L^2([0,1])$ is an unknown function, $\sigma > 0$ is a known constant, and W is a standard Brownian motion. This rewrites in integral form as

$$Y(t) = \int_0^t f(s)\,ds + \frac{\sigma}{\sqrt{n}} W(t), \quad t \in [0,1],$$

and defines a stochastic process $Y := (Y(t))_{t \in [0,1]}$ which is assumed observed. Since Y is observed, we naturally observe the random variable $\int_0^1 g(t)\,dY(t)$ for any $g \in L^2([0,1])$. From the differential equation defining Y, we directly have for any $g \in L^2([0,1])$ that

$$\int_0^1 g(t)\,dY(t) = \int_0^1 g(t) f(t)\,dt + \frac{\sigma}{\sqrt{n}} \int_0^1 g(t)\,dW(t).$$

If we take an orthonormal basis $(t \mapsto \varphi_i(t))_{i \in \mathbb{N}}$ of $L^2([0,1])$ and define

$$Y_i = \int_0^1 \varphi_i(t)\,dY(t), \quad \theta_i = \int_0^1 \varphi_i(t) f(t)\,dt, \quad Z_i = \int_0^1 \varphi_i(t) dW(t),$$

then the previous equation yields

$$Y_i = \theta_i + \frac{\sigma}{\sqrt{n}} Z_i,$$

where θ_i is deterministic, $Z_i \sim N(0, \|\varphi_i\|_2^2) = N(0,1)$, and $\mathrm{cov}(Z_i, Z_j) = \delta_{i,j}$ (the two last properties following from the properties of the Brownian motion). It follows that the Gaussian white noise model is observationally equivalent to the Gaussian sequence model. We can generalize the Gaussian white noise model as we did with the Gaussian sequence model by considering

$$dY(t) = f(t)\,dt + \sigma \zeta_n dW(t), \quad t \in [0,1].$$

In this case, the connection between the two models generalize without modification.

3 Links to (Gaussian) nonparametric regression

A **nonparametric regression** is a model of the form

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where Y_i is an observed real random variable, $f\colon [0,1] \to \mathbb{R}$ is some unknown function (called the **regression function**), $x_i \in [0,1]$ is some observed deterministic value, and ε_i is some unobserved random variable. The model naturally generalizes to any compact interval $[a,b]$. Since the x_i are deterministic, the model is said to be of **fixed design**. If it is assumed in the fixed design that $x_i = i/n$, then the design is said to be **equally spaced**. (This can be understood as sampling the unknown function on an equally spaced grid of [0,1] that grows dense in [0,1] as $n \to \infty$ – Gine&Nickl MFSIDSM p.6). It is also possible to generalize the model by taking instead of observed deterministic x_i observed random variables X_i (which reduces to the latter by taking $X_i = x_i$ a.s.). In this case, the model is said to be of **random design** (and "[o]ne can then either proceed to argue conditionally on the realisations $X_i = x_i$, or one takes this randomness explicitly into account by making probability statements under the law[s of the X_i and ε_i] simultaneously." Gine&Nickl MFSIDSM p.6). If the ε_i are assumed to be identically distributed according to $N(0,\sigma^2)$ for some $\sigma > 0$, then the nonparametric regression is said to be **Gaussian**. (We could also write $\varepsilon_i = \sigma Z_i$ where $Z_i \sim N(0,1)$ to match the previous notations). It is generally assumed that the ε_i are not only identically distributed but also independent. (We will always assume it in the definition of Gaussian in what follows). The objective of nonparametric regression is to recover the unknown regression function f. It is a nonparametric problem since f is infinite-dimensional. For the task to make sense (that is, estimating an infinite-dimensional object by only sampling finitely many values), it is necessary to restrict "slightly" the class $\mathscr{F}$ in which f belongs (but not up to parametricity, that is, a finite-dimensional class $\mathscr{F}$). Examples of such classes include the set of continuous functions on $[0,1]$, the set of convex functions on $[0,1]$, etc.

The **Gaussian nonparametric regression with fixed equally spaced design** on $[0,1]$ is thus given by

$$Y_i = f\left(\frac{i}{n}\right) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0,\sigma^2), \quad i = 1,\dots,n,$$

where $f\colon [0,1] \to \mathbb{R}$ belongs in some class $\mathscr{F}$ of functions. We now informally show that this model is very close (and asymptotically equivalent for a rigorously defined notion of equivalence as that of Le Cam) to the Gaussian sequence model and hence also to the Gaussian white noise model. Let $(\varphi_i)_{i\in\mathbb{N}}$ be an orthonormal basis of $L^2([0,1])$ and define

$$\gamma_i = \frac{1}{n}\sum_{k=1}^{n} Y_k \varphi_i\left(\frac{k}{n}\right), \quad f_i = \frac{1}{n}\sum_{k=1}^{n} f\left(\frac{k}{n}\right)\varphi_i\left(\frac{k}{n}\right), \quad \zeta_i = \frac{1}{\sqrt{n}}\sum_{k=1}^{n} Z_i \varphi_j\left(\frac{k}{n}\right).$$

Then the model equation directly yields

$$\gamma_i = f_i + \frac{\sigma}{\sqrt{n}}\zeta_i, \quad i = 1\dots,n,$$

which is seen to be a finite approximation version of the Gaussian sequence model as derived from the Gaussian white noise model.

References

Gine, E., and R. Nickl (2021): *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press.

Ingster, Y. I., and I. A. Suslina (2003): *Nonparametric goodness-of-fit testing under gaussian models*. Springer.

Johnstone, I. M. (2019): "Gaussian estimation: Sequence and wavelet models," *Unpublished lecture notes*.

Tsybakov, A. B. (2008): *Introduction to nonparametric estimation*. Springer.

Wasserman, L. (2020): "Lecture Notes 22/36-705," *Lecture notes for Statistics 36-705 at CMU*.