Universiteit
Leiden
The Netherlands

Bayesian inference for Gaussian models: Inverse problems and evolution equations

Yan, D.

Citation
Yan, D. (2020, March 3). *Bayesian inference for Gaussian models: Inverse problems and evolution equations*. Retrieved from https://hdl.handle.net/1887/86070

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/86070

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

The handle http://hdl.handle.net/1887/86070 holds various files of this Leiden University dissertation.

Author: Yan, D.
Title: Bayesian inference for Gaussian models: Inverse problems and evolution equations
Issue Date: 2020-03-03

Chapter 4

Bayesian Nonparametrics for Gaussian Linear Models

In this chapter, we survey some basic facts of Bayesian nonparametrics, retrieved from [35]. In addition, we present a general contraction results for the Gaussian linear model with a transformed drift.

4.1 Bayesian Nonparametrics

4.1.1 Bayes' Rule

As mentioned in Section 1.3, Bayes' rule on infinite-dimensional spaces requires special care on measurability concerns. In this section, we sketch Bayes' rule in infinite-dimensional spaces. For more details, see Section 1.3 in [35].

The Bayesian procedure can be described in the following steps. First, the statistician seeks for a *prior distribution* Π to apply on the parameter space $(\Theta, \mathscr{S})$, where $\mathscr{S}$ is a σ-algebra. Then, given θ, each element $\mathbb{P}_\theta$ in the statistical model $\{\mathbb{P}_\theta : \theta \in \Theta\}$ becomes a regular conditional distribution on the sample space $(\mathbb{X}, \mathscr{X})$. That means, the mapping $\Theta \times \mathscr{X} \ni (\theta, A) \mapsto \mathbb{P}_\theta(A) \in [0,1]$ is a *(probability) kernel* from Θ to $\mathbb{X}$, i.e.

(i) with fixed θ, $A \mapsto \mathbb{P}_\theta(A)$ is a probability measure on $\mathbb{X}$,

(ii) and with fixed A, $\theta \mapsto \mathbb{P}_\theta(A)$ is $\mathscr{S}$-measurable.

Therefore, the pair (X, θ) has a well-defined joint distribution

$$\mathbb{P}(X \in A, \theta \in B) = \int_B \mathbb{P}_\theta(A)\, d\Pi(\theta)$$

on the product space $(\mathbb{X} \times \Theta, \mathscr{X} \times \mathscr{S})$. Consequently, the *(Bayesian) marginal distribution* of X is given by

$$\mathbb{P}(A) = \int \mathbb{P}_\theta(A)\, d\Pi(\theta), \quad A \in \mathscr{X},$$

and the *posterior distribution*, the conditional distribution of θ given X, is

$$\Pi(B \mid X) = \mathbb{P}(\theta \in B \mid X), \quad B \in \mathscr{S}.$$

Remark 4.1. The posterior distribution is always well-defined, by Kolmogorov's definition. That means, $\Pi(\theta \mid X)$ is a measurable function of X such that, for a fixed $B \in \mathscr{S}$ and any $A \in \mathscr{X}$, $\mathbb{E}[\Pi(\theta \mid X)\, \mathbb{1}\{X \in A\}] = \mathbb{P}(X \in A, \theta \in B)$. However, to obtain a *regular* version of the conditional distribution, the size of the space $(\Theta, \mathscr{S})$ cannot be too big. One sufficient condition is that Θ is a Polish space, i.e. a complete *separable* metric space, and $\mathscr{S}$ is the Borel σ-algebra.

It is noteworthy that the posterior distribution $\Pi(\cdot \mid X)$ is unique up to null sets under the Bayesian marginal distribution. For a faithful Bayesian, this does not cause any problems, since it is believed that the Bayesian marginal distribution generates the data. However, in the case that the Bayesian framework is treated as an inference method, a 'true' distribution $\mathbb{P}_0$ generating the data X is not necessary identical to the Bayesian marginal distribution, and hence they do not have the same null sets. This phenomenon leads to indefiniteness of the posterior distribution. The pathological circumstance above is related to *misspecification* and can be excluded by the following condition,

$$\mathbb{P}_0 \ll \int \mathbb{P}_\theta \; d\Pi(\theta).$$

When the statistical model $\mathscr{E} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ is dominated by a σ-finite measure μ, the Radon-Nikodym derivative provides densities p_θ related to the dominating measure μ such that $(x, \theta) \mapsto p_\theta(x)$ is measurable. Then, the density version of *Bayes' formula* is given by

$$\Pi(B \mid X) = \frac{\int_B p_\theta(X)\, d\Pi(\theta)}{\int p_\theta(X)\, d\Pi(\theta)}. \tag{4.1}$$

4.1.2 Bayesian Asymptotics

For a given Bayesian inferential procedure, i.e. estimating a parameter using the posterior distribution generated by Bayes' rule and a prior on the parameter space, its statistical performance can be evaluated in the asymptotic framework. Specifically, the following frequentist concepts are used to characterise posteriors.

For a sequence of experiments $\mathscr{E} = \{\mathbb{X}^{(n)}, \mathscr{X}^{(n)}, \mathbb{P}_\theta^{(n)} : \theta \in \Theta\}$, consider a prior Π on the Borel-algebra $\mathscr{B}(\Theta)$, and fix a version $\Pi_n(\cdot \mid X^{(n)})$ of its posterior distribution, i.e. any given choice of a regular condition distribution of θ given $X^{(n)}$ (see Remark 4.1), which is referred to as *the* posterior.

The asymptotic consistency for Bayesian procedures is defined as follows.

Definition 4.2 (Consistency)**.** The posterior $\Pi_n(\cdot \mid X^{(n)})$ is *(weakly) consistent* at $\theta_0 \in \Theta$ if, for every neighbourhood U of θ_0, in $\mathbb{P}_{\theta_0}^{(n)}$ probability,

$$\Pi_n(U^c \mid X^{(n)}) \to 0, \quad \text{as } n \to \infty,$$

and it is *strongly consistent* if the convergence above is $\mathbb{P}_{\theta_0}^{(n)}$-almost surely.

The concept of contraction is useful for quantifying the speed that the posterior concentrates around the truth.

Definition 4.3 (Contraction rate)**.** Suppose that the parameter space Θ is equipped with a pseudo-metric $d(\cdot, \cdot)$. A sequence ε_n is a *posterior contraction rate* at the parameter θ_0 if

$$\Pi_n\Big(\theta : d(\theta, \theta_0) \geq M_n \varepsilon_n \mid X^{(n)}\Big) \to 0, \quad \text{as } n \to \infty,$$

in $\mathbb{P}_{\theta_0}^{(n)}$, for arbitrarily slow $M_n \to \infty$.

Definition 4.3 requires some clarifications. First, the sequence ε_n is a rate rather than *the* rate, since any sequence decreasing slower also satisfies the definition. Certainly the fastest decaying sequence ε_n is of interest, but it may not be tractable in general. Therefore, ε_n is actually an *upper bound* for a targeted rate,but we will refer to it as *the* rate, with an abuse of terminology. In addition, due to the inaccessibility of minimal rates, we are satisfied with the result that the contraction rates are comparable (matching or close to) some benchmarks, for which a frequently used example is the optimal rate in minimax sense.

For statistical models in infinite-dimensional spaces, it is often the case that the unbounded sequence M_n can be replaced by a sufficiently large constant M, which gives a slightly stronger result. On the other hand, for parametric models, the unboundedness is required in order to obtain the regular rate $n^{-1/2}$. In this thesis, in all chapters we consider the strong version of Definition 4.3, i.e. M being a fixed constant, except in Chapter 7.

Some point estimators naturally emerge from the posterior distribution. An example is the centre of the smallest ball containing at least posterior mass $1/2$. Another one is the posterior mean. For further details, see Section 8.1 in [35].

In the Bayesian approach, credible sets are used to quantify uncertainty, whose counterpart in frequentist framework is confidence regions.

Definition 4.4 (Credible set)**.** Fix a constant $1 - \gamma \in (0, 1)$. A subset $\mathcal{C}_n(X^{(n)})$ of Θ is a *credible set* for θ of credibility level $1 - \gamma$ if

$$\Pi_n\Big(\theta \in \mathcal{C}_n(X^{(n)}) \mid X^{(n)}\Big) \geq 1 - \gamma,$$

and the *frequentist coverage* of credible sets is given by

$$\mathbb{P}_\theta\Big(\theta \in \mathcal{C}_n(X^{(n)})\Big).$$

It is logical to consider the sets capturing most of the posterior mass while retaining a reasonable size at the same time. For example, if the posterior is unimodal, a ball centred at its mode can be considered as a credible set. Credibility is not the main focus of this thesis, but we will touch it in Chapter 7.

4.2 Gaussian Linear Models

In Section 1.2, we have already mentioned the following model,

$$X^{(n)} = \mathcal{A}^{(n)}\theta + \xi^{(n)}.$$

In this section, we formally define the model, using the elements introduced in Chapters 2 and 3.

There are two components in the model: the transformed signal $\mathcal{A}^{(n)}\theta$ and the noise $\xi^{(n)}$, and our interest is the original signal θ. Different combinations of $\mathcal{A}^{(n)}$ and $\xi^{(n)}$ leads to separate models. The choice of $\mathcal{A}$ highly depends on the problem under investigation. In Part II, $\mathcal{A}$ is a forward mapping with smoothing property, related to inverse problems, and in Part III, two items from the solution mapping of evolution equations are represented as $\mathcal{A}$. In the present chapter, we only assume that $\mathcal{A} : \Theta \to G$ is a bounded linear operator, from the parameter space Θ to another Hilbert space G, both of which are separable, and our focus is on the interpretation of noise. We are going to discuss two types of noise related to different observation schemes.

4.2.1 Continuous Observation

We consider the following Gaussian linear model with continuous observations. For $n \in \mathbb{N}$, let

$$\mathcal{A}^{(n)} = \mathcal{A} \quad \text{and} \quad \xi^{(n)} = \frac{1}{\sqrt{n}}\xi,$$

i.e.

$$X^{(n)} = \mathcal{A}\theta + \frac{1}{\sqrt{n}}\xi. \tag{4.2}$$

The forward operator is independent of the experiment sequence, while the noise $\xi^{(n)}$ is a fixed Gaussian element ξ scaled by $1/\sqrt{n}$, representing more information collected as $n \to \infty$. The observation is called *continuous* if we record the entire signal $\mathcal{A}\theta$ in G with noise ξ. Using the duality structure of Hilbert space G, the complete signal trajectory $\mathcal{A}\theta$ is equivalent to $(\langle \mathcal{A}\theta, g\rangle_G : g \in G)$. Concerning the noise, the conventional framework in statistics is Gaussian processes indexed by space G, which will be illustrated below. We conclude the discussion with a formula of the Kullback-Leibler distance between two observations with different signals $\mathcal{A}\theta$.

In our measurement model the observation $X^{(n)}$ will be a stochastic process $(X^{(n)}(w) : w \in G)$ such that

$$X^{(n)}(w) = \langle \mathcal{A}\theta, w\rangle_G + \frac{1}{\sqrt{n}}\xi(w), \qquad w \in G, \tag{4.3}$$

where $\xi = (\xi(w) : w \in G)$ is a Gaussian process defined as follows. Let $\mathcal{Q} : G \to G$ be a bounded self-adjoint positive-definite operator. With the inner product $\langle \cdot, \cdot \rangle_G$ and norm $\|\cdot\|_G$ on G, we consider the noise ξ to be a Gaussian process indexed

by the Hilbert space G such that, for every $h, g \in G$, $\xi(h)$ is Gaussian with zero mean, i.e. $\mathbb{E}\,\xi(h) = 0$, and

$$\mathbb{E}\,\xi(h)\xi(g) = \langle h, \mathcal{Q}g\rangle_G. \tag{4.4}$$

By abuse of notation, we also call $\mathcal{Q}$ the covariance operator of the Gaussian process ξ.

The processes $X^{(n)}$ and ξ are viewed as measurable maps in the *sample space* $\mathbb{R}^G$, with its product σ-field. Statistical sufficiency considerations show that the observation can also be reduced to the vector $(X^{(n)}(w_1), X^{(n)}(w_2), \ldots)$, which takes values in the sample space $\mathbb{R}^{\mathbb{N}}$, for any orthonormal basis $(w_i)_{i\in\mathbb{N}}$ of G. The coordinates $X^{(n)}(w_i)$ of this vector are random variables with normal distributions such that

$$\mathbb{E}\,X^{(n)}(w_i) = \langle \mathcal{A}\theta, w_i\rangle_G, \quad \mathbb{E}\Big(X^{(n)}(w_i)\Big)^2 = \frac{1}{n}\|\mathcal{Q}^{1/2}w_i\|_G^2,$$

with covariance $\mathbb{C}\mathrm{ov}\big(X^{(n)}(w_i), X^{(n)}(w_j)\big) = \langle w_i, \mathcal{Q}w_j\rangle_G$ given by (4.4). If the basis $(w_i)_{i\in\mathbb{N}}$ is also orthogonal with respect to the inner product induced by $\mathcal{Q}$, i.e. $\langle w_i, \mathcal{Q}w_j\rangle_G = 0$ if $i \neq j$, then the variables $\xi(w_1), \xi(w_2), \ldots$ are stochastically independent normal variables. In this case, $(X^{(n)}(w_1), X^{(n)}(w_2), \ldots)$ is known as the *Gaussian sequence model* in statistics, albeit presently the 'drift function' $\mathcal{A}\theta$ involves the operator $\mathcal{A}$. See [11, 50] and references therein. In the rest of this thesis, we will consider the following cases.

(i) In Part II, $\mathcal{Q} = \mathrm{id}$ on G. ξ is an isonormal process on G and cannot be realised as a proper element of G. However, (4.3) makes perfect sense and (4.4) corresponds to $\mathbb{E}\,\xi(h)\xi(g) = \langle h, g\rangle_G$, the defining equation of isonormal process. We take $\mathbb{G} = \mathrm{id}(G)$.

(ii) In Part III, $\mathcal{Q} \neq \mathrm{id}$ on G. The covariance structure of ξ is characterised by $\mathcal{Q}$. If $\mathcal{Q}$ is of trace class, then ξ induces a centred Gaussian measure on G, whose RKHS is $\mathbb{G} = \mathcal{Q}^{1/2}(G)$ equipped with norm $|h|_{\mathbb{G}} = \|\mathcal{Q}^{-1/2}h\|_G$. Consequently, (4.2) defines a Borel mapping $X^{(n)}$ into G, and the interpretation $X^{(n)}(g) = \langle X^{(n)}, g\rangle_G$ leads to (4.3). Otherwise, if $\mathcal{Q}$ is not of trace class, the same argument from the white noise case applies.

In both cases, the law of $X^{(n)}$ is dominated by $\xi/\sqrt{n}$ is equivalent to $\mathcal{A}\theta \in \mathcal{Q}^{1/2}(G) = \mathbb{G}$. So we assume $\mathrm{Ran}\,\mathcal{A} \subset \mathrm{Ran}(\mathcal{Q}^{1/2})$. Moreover, the Kullback-Leibler distance is given in the lemmas below. For the proper Gaussian case, it is a direct consequence of Lemma 3.20.

Lemma 4.5. *Let ξ be a proper centred Gaussian on a Hilbert space G with $\mathbb{G}$ being the RKHS. Let $\gamma_{g_1}, \gamma_{g_2}$ be the law of*

$$X_1 = g_1 + \xi \quad \textit{and} \quad X_2 = g_2 + \xi$$

respectively. Then, we have

$$\mathbb{E}_{g_1}\left[\log\frac{d\gamma_{g_1}}{d\gamma_{g_2}}\right] = \frac{1}{2}\|g_1 - g_2\|_{\mathbb{G}}^2 \quad \textit{and} \quad \mathbb{V}\mathrm{ar}_{g_1}\left[\log\frac{d\gamma_{g_1}}{d\gamma_{g_2}}\right] = \|g_1 - g_2\|_{\mathbb{G}}^2,$$

where the subscript g_1 denotes that the integrals are calculated with respect to measure γ_{g_1}.

For the white noise case, it requires more elaboration. We consider the observations in the sequence model.

Lemma 4.6. *For $\theta = (\theta_1, \theta_2, \ldots)$ let P_θ be the distribution of the random element $(Z_1 + \theta_1, Z_2 + \theta_2, \ldots)$ in $\mathbb{R}^\infty$ for $Z_1, Z_2 \ldots$ i.i.d. mean-zero normal variables with variance σ^2. If $\theta \in \ell^2$, then P_θ is absolutely continuous relative to P_0 with log likelihood*

$$\log \frac{dP_\theta}{dP_0}(X_1, X_2, \ldots) = \frac{1}{\sigma^2} \sum_{i=1}^\infty \theta_i X_i - \frac{1}{2\sigma^2} \sum_{i=1}^\infty \theta_i^2,$$

where the first series converges almost surely and in second mean. The expectation and variance of this variable are

$$\mathbb{E}_\theta \left[\log \frac{dP_\theta}{dP_0} \right] = \frac{1}{2\sigma^2} \sum_{i=1}^\infty \theta_i^2 \quad \textit{and} \quad \mathbb{V}\mathrm{ar}_\theta \left[\log \frac{dP_\theta}{dP_0} \right] = \frac{1}{\sigma^2} \sum_{i=1}^\infty \theta_i^2.$$

Proof. That the series converges in L^2 is clear from the fact that $\theta \in \ell^2$; the almost sure convergence next follows from the Itô-Nisio theorem. The expectation and variance of the right side are easy to compute as limits.

Write Σ_∞ for the right side of the display, and Σ_n for the expression obtained by replacing the infinite sums by the sums from 1 to n. Thus $\Sigma_n \to \Sigma_\infty$ almost surely. Since $\mathbb{E}_0 e^{2\Sigma_n} = e^{\sum_{i=1}^n \theta_i^2/\sigma^2}$ is uniformly bounded in n, it follows that e^{Σ_n} is uniformly integrable and hence converges in mean to e^{Σ_∞}. In particular, the mean of the latter variable is 1, the mean of the former variables.

It follows that the Borel measure on $\mathbb{R}^\infty$ defined by $B \mapsto \mathbb{E}_0 1_B(X) e^{\Sigma_\infty}$ is a probability measure. For every Borel set B it is the limit of $\mathbb{E}_0 1_B(X) e^{\Sigma_n}$, which is $P_\theta(B)$ if B depends only on the first n coordinates, as e^{Σ_n} is the density of the distribution of $(Z_1 + \theta_1, \ldots, Z_n + \theta_n)$ with respect to its distribution at $\theta = 0$. Since the Borel σ-field on $\mathbb{R}^\infty$ is generated by the algebra of all cylinder sets, it follows that P_θ and the measure $B \mapsto \mathbb{E}_0 1_B(X) e^{\Sigma_\infty}$ agree. □

Recall that a bounded operator $\mathcal{Q} : H \to H$ is called *diagonalisable*, if with an orthonormal basis $\{\varphi_k\}_{k \in \mathbb{N}}$ for H, for all $f \in H$,

$$\mathcal{Q} f = \sum_{k \in \mathbb{N}} q_k f_k \varphi_k \quad \text{with } f_k = \langle f, \varphi_k \rangle_H.$$

The last result can be easily extended to the case of diagonalisable operators.

Corollary 4.7. *Let Z_k in Lemma 4.6 be independent mean-zero normal variables with variance σ_k^2 such that $\sup_k \sigma_k < \infty$. Then, the same result holds with*

$$\log \frac{dP_\theta}{dP_0}(X_1, X_2, \ldots) = \sum_{i=1}^\infty \frac{\theta_i}{\sigma_i^2} X_i - \frac{1}{2} \sum_{i=1}^\infty \frac{\theta_i^2}{\sigma_i^2},$$

and

$$\mathbb{E}_\theta \left[\log \frac{dP_\theta}{dP_0} \right] = \frac{1}{2} \sum_{i=1}^\infty \frac{\theta_i^2}{\sigma_i^2} \quad \textit{and} \quad \mathbb{V}\mathrm{ar}_\theta \left[\log \frac{dP_\theta}{dP_0} \right] = \sum_{i=1}^\infty \frac{\theta_i^2}{\sigma_i^2}.$$

4.2.2 Discrete Observation

In practice, it is rare that complete information can be acquired from measurements, and only *partial* observations can be gathered. It is often designated by a natural number n, representing the amount of observations having been made. We consider the following type of partial observation. Let G be the space $L^2(\mathfrak{D})$ of square integrable functions on a bounded domain $\mathfrak{D} \subset \mathbb{R}^d$ with $d \in \mathbb{N}$. Instead of observing the entire trajectory of $\mathcal{A}\theta$ in G, we observe the process at the *design points*, a set $\mathfrak{D}_n = \{x_i\}_{i \le n} \subset \mathfrak{D}$. To formalise the description, let

$$\mathcal{A}^{(n)} = \mathcal{E}^{(n)}\mathcal{A}, \quad \xi^{(n)} = (z_i)_{i \le n},$$

where $\mathcal{E}^{(n)} : G \to \mathbb{R}^n$ is the evaluation operator at design points, and $(z_i)_{i \le n}$ is the standard Gaussian in $\mathbb{R}^n$. The observation $X^{(n)}$ is a random vector $(X_i)_{i \le n}$ in $\mathbb{R}^n$, with the coordinates given by

$$X_i = (\mathcal{A}\theta)(x_i) + z_i, \quad i = 1, \cdots, n,$$

where z_i are i.i.d standard Gaussian variables.

In literature (e.g. [42]), the majority of discrete observations refers to the aforementioned type. While, less frequently, it may address the continuous observations being truncated, e.g. only the first n entries of the infinite vector $(X^{(n)}(w_k))_{k \in \mathbb{N}}$. We do not follow this convention, but it is often equivalent to the discrete observations at design points, with regularity assumptions on the signal $\mathcal{A}\theta$. For example, in Section 8.1, we will also present a concrete construction that the observation at design points implies the truncated continuous observation.

Connection to white noise model

We are going to give some heuristics to link the regression to the white noise model. Since only the observation is relevant, the shorthand notation $g = \mathcal{A}\theta$ is used. We assume that there exists a partition $\{\mathfrak{D}_i : 1 \le i \le n\}$ of $\mathfrak{D}$ such that the subdomains are mutually disjoint and of equal volume $\operatorname{vol}\mathfrak{D}_i \simeq 1/n$, and furthermore each subdomain $\mathfrak{D}_i$ contains only one design point x_i.

The observations in white noise can be reduced to the projections on a orthonormal basis $\{v_k\}_{k \in \mathbb{N}}$ of $L^2(\mathfrak{D})$. If the basis functions are continuous as well as the drift function g, the projection can be approximated by the Riemann summation,

$$\int_{\mathfrak{D}} g(x)v(x)\,dx \approx \sum_{i=1}^{n} g(x_i)v(x_i)\operatorname{vol}\mathfrak{D}_i.$$

Recall $\operatorname{vol}\mathfrak{D}_i \simeq 1/n$. By plugging in the noisy observations from regression, the noise is

$$\sum_{i=1}^{n} z_i \frac{v(x_i)}{n} \sim \mathscr{N}_{\mathbb{R}}\left(0, \frac{1}{n}\sum_{i=1}^{n} v(x_i)^2 \operatorname{vol}\mathfrak{D}_i\right).$$

On the other hand, from Example 3.31,

$$\frac{1}{\sqrt{n}}\xi(v) = \frac{1}{\sqrt{n}}\int_{\mathfrak{D}} \xi(t)v(t)\,dt \sim \mathscr{N}_{\mathbb{R}}\left(0, \frac{1}{n}\int_{\mathfrak{D}} v(x)^2\,dt\right).$$

Both of the noises are centred and their variances are asymptotically equal.

Remark 4.8 (Continuity)**.** It is noteworthy that the assumption on the continuity of both drift g and the test function v is crucial. As we will see in the subsequent Chapter 7 and Chapter 8, the drift $\mathcal{A}f$ needs to possess certain smoothness, which corresponds to the minimal requirement on securing continuity.

The heuristics above shows that the regression model converges to the continuous model in a certain sense, while more data are gathered. However, conversely, the point evaluations of white noise model does not lead to a regression model. This is because point evaluation is only defined for continuous functions, but the while noise ξ is only distribution-valued (see Example 3.31). One cannot talk about the values of ξ evaluated at points.

4.3 Bayesian Contraction for Gaussian Linear Models

In this section we present a general theorem on the posterior contraction for linear problems in the following form. First we introduce the smoothness class. Let $(\Theta, \langle\cdot,\cdot\rangle)$ be a separable Hilbert space. Given a set $\mathfrak{S}$ in a finite dimensional ordered space (e.g. $\mathbb{R}^m$, $m \in \mathbb{N}$), assume that there exists a smoothness scale $\{\Theta_s, \langle\cdot,\cdot\rangle_s\}_{s\in\mathfrak{S}}$ with $\Theta_0 = \Theta$, such that for $s < t$, the induced norm $\|\cdot\|_t$ is strictly stronger than $\|\cdot\|_s$ and $\Theta_t \subset \Theta_s$. We consider the observation scheme as introduced in Section 4.2.1, a transform $\mathcal{A}\theta$ of θ in G with Gaussian process ξ indexed by G,

$$X^{(n)} = \mathcal{A}\theta + \frac{1}{\sqrt{n}}\xi. \tag{4.5}$$

Let ξ be a centred Gaussian process indexed by a separable Hilbert space G with covariance operator $\mathcal{Q}$. Assume that $\mathcal{A} : \Theta \to G$ is a bounded linear operator satisfying

$$\operatorname{Ran}\mathcal{A} \subset \mathbb{H} := \mathcal{Q}^{1/2}(G),$$

where $\mathbb{H}$ is a Hilbert space equipped with the induced inner product (see Lemma A.11),

$$\langle\cdot,\cdot\rangle_{\mathbb{H}} := \langle\mathcal{Q}^{-1/2}\cdot, \mathcal{Q}^{-1/2}\cdot\rangle_G.$$

We will only consider the *non-degenerate* case, i.e. $\mathbb{H}$ is dense in G. To link the spaces G and Θ, we introduce the following assumption.

Assumption 4.9. The family $\{\mathcal{R}_n\}_{n\in\mathbb{N}}$ of linear reconstruction operators $\mathcal{R}_n : \mathbb{H} \to \Theta$ satisfies the following properties.

(i) With $j_n \to \infty$, let $\{W_{j_n}\}_{n\in\mathbb{N}}$ be a sequence of subspaces of G such that, for all $n \in \mathbb{N}$,

$$\dim W_{j_n} = j_n, \quad W_{j_n} \subset W_{j_{n+1}}, \quad W_{j_n} \in \mathbb{H},$$

and the kernel (i.e. the null space) of $\mathcal{R}_n$ is $W_{j_n}^{\perp}$.

(ii) Let ρ_n be a monotonically nondecreasing sequence which may go to infinity as $n \to \infty$. $\mathcal{R}_n$ satisfies

$$\|\mathcal{R}_n\|_{G;\Theta} \simeq \rho_n. \tag{4.6}$$

(iii) For $s \in \mathfrak{S}$, let $\{\delta(j_n, s)\}_{s\in\mathcal{I}}$ be a family of monotonically decreasing functions such that $\delta(j_n, s) \downarrow 0$ as $n \to \infty$, for all $s \in \mathfrak{S}$. We assume

$$\|\mathcal{R}_n\mathcal{A}\theta - \theta\|_\Theta \leq \delta(j_n, s)\|\theta\|_s \tag{4.7}$$

holds, for $\theta \in \Theta_s$, $s \in \mathfrak{S}$.

We form the posterior distribution $\Pi_n(\cdot \mid X^{(n)})$ as in (4.1), given a prior Π_n on the space $\Theta = \Theta_0$ and an observation $X^{(n)}$, whose conditional distribution given θ is determined by the model (4.5). We study this random distribution under the assumption that $X^{(n)}$ follows the model (4.5) for a given 'true' function $\theta = \theta_0$, which we assume to be an element of Θ_β in a given smoothness class $(\Theta_s)_{s\in\mathfrak{S}}$.

The result is based on an extension of the testing approach of [35] to the parameter inference of (4.5). The recovery problem is handled with the help of the reconstruction family from Assumption 4.9. We note that the reconstruction family only appears as a tool to state and derive a posterior contraction rate. In our context it does not enter into the solution of the linear problem, which is achieved through the Bayesian method.

Clearly the reconstructed signal $\theta^{(n)} = \mathcal{R}_n\mathcal{A}\theta$ is an approximation to θ, which will be better for increasing n, but increasingly complex, when ρ_n is unbounded. The following theorem uses ρ_n that balances approximation to complexity, where the complexity is implicitly determined by a testing criterion.

Theorem 4.10. *Assume $\theta_0 \in \Theta_\beta$ with $\beta \in \mathbb{R}^m_+$ such that $\mathcal{A}\Theta_\beta \subset \mathbb{H}$.*

For $\varepsilon_n \downarrow 0$ such that $n\varepsilon_n^2 \to \infty$, suppose Assumption 4.9 holds with

$$j_n \leq cn\varepsilon_n^2, \tag{4.8}$$

where c is a positive constant. In addition, for $\eta_n \geq \varepsilon_n$, assume

$$\eta_n \geq \rho_n\varepsilon_n, \tag{4.9}$$

$$\eta_n \geq \delta(j_n, \beta). \tag{4.10}$$

Let $\theta^{(n)}$ denote a reconstruction estimator $\mathcal{R}_n\mathcal{A}\theta$ to θ. Consider prior probability distributions Π on Θ satisfying

$$\Pi(\theta : \mathcal{A}\theta \in \mathbb{H}) = 1, \tag{4.11}$$

$$\Pi(\theta : \|\mathcal{A}\theta - \mathcal{A}\theta_0\|_\mathbb{H} < \varepsilon_n) \geq e^{-n\varepsilon_n^2}, \tag{4.12}$$

$$\Pi(\theta : \|\theta^{(n)} - \theta\|_0 > \eta_n) \leq e^{-4n\varepsilon_n^2}. \tag{4.13}$$

Then the posterior distribution in the model (4.5) contracts at the rate η_n at θ_0, i.e. for a sufficiently large constant M we have $\Pi_n(\theta : \|\theta-\theta\|_0 > M\eta_n \mid X^{(n)}) \to 0$, in probability under the law of $X^{(n)}$ given by (4.5) with $\theta = \theta_0$.

The conditions in Theorem 4.10 deserve some explanations.

The desirable structure of priors is characterised by eqs. (4.11) to (4.13). Since we only consider dominated models in Gaussian noise, (4.11) assures the validity of the density version (4.1) of Bayes' formula. (4.12) is the *prior mass* condition,

that is, a set around the truth should have sufficient probability mass at most exponentially decay with $n\varepsilon_n^2$. The last condition (4.13) states a requirement on the prior related to the testing approach via the constructed signals. It says that the prior mass on a set, whose elements, even without noise, cannot be accurately reconstructed quantified by η_n, should have prior mass upper bounded by a negative exponential of $n\varepsilon_n^2$.

The final contraction rate is determined with the constraints eqs. (4.8) to (4.10). The first (4.8) can be interpreted as a consideration that the reconstruction should be capped by the available information, indicated by $n\varepsilon_n^2$. The second constraint (4.9) shows the cost to recover the original signal θ from the transform $\mathcal{A}$, and the last one (4.10) simply states the fact that the rate cannot be faster than the recovery rate for noiseless signals.

We conclude this section with proving the main theorem of this section.

Proof of Theorem 4.10. Using Lemma 4.5, Lemma 4.6, and Corollary 4.7 from Section 4.2.1, the Kullback-Leibler divergence and variation between the distributions of $X^{(n)}$ under two functions θ and θ_0 are

$$n\|\mathcal{A}\theta - \mathcal{A}\theta_0\|_{\mathbb{H}}^2/2 \quad \text{and} \quad n\|\mathcal{A}\theta - \mathcal{A}\theta_0\|_{\mathbb{H}}^2.$$

Therefore the neighbourhoods $B_{n,2}(\theta_0, \varepsilon)$ in (8.19) of [35] contain the ball $\{\theta \in \Theta : \|\mathcal{A}\theta - \mathcal{A}\theta_0\|_{\mathbb{H}} \leq \varepsilon\}$. By assumption (4.12) this has prior mass at least $e^{-n\varepsilon_n^2}$.

Because the quotient of the left sides of (4.12) and (4.13) is $o(e^{-2n\varepsilon_n^2})$, the posterior probability of the set $\{f : \|\theta^{(n)} - \theta\|_0 > \eta_n\}$ tends to zero, by Theorem 8.20 in [35].

By a variation of Theorem 8.22 in [35] it is now sufficient to show the existence of tests τ_n such that, for some $M > 0$,

$$\mathbb{P}_{\theta_0}^{(n)} \tau_n \to 0, \qquad \sup_{\substack{\theta : \|\theta - \theta_0\|_0 > M\eta_n, \\ \|\theta^{(n)} - \theta\|_0 \leq \eta_n}} \mathbb{P}_{\theta}^{(n)}(1 - \tau_n) \leq e^{-4n\varepsilon_n^2}.$$

Indeed, in the case that the prior mass condition (8.20) in Theorem 8.22 of [35] can be strengthened to (8.22), as is the case in our setup in view of (4.12), it suffices to verify (8.24) only for a single value of j. Furthermore, we can apply Theorem 8.22 with the metrics $d_n(x, y) = \|x - y\|_0 \varepsilon_n / \eta_n$ in order to reduce the restriction $d_n(\theta, \theta_{n,0}) > M\varepsilon_n$ to $\|\theta - \theta_0\|_0 > M\eta_n$.

Let $\{e_k\}_{k \leq j_n}$ be an G-orthonormal basis of W_{j_n}, and denote the G-orthogonal projection onto W_{j_n} by

$$\mathcal{P}_{j_n} : G \to W_{j_n} \subset \mathbb{H}.$$

By slight abuse of notation, define the projection $\mathcal{P}_{j_n}$ of process ξ onto W_{j_n} by

$$\mathcal{P}_{j_n} \xi := \sum_{i \leq j_n} \xi(e_k) e_k,$$

where $\xi(e_k)$ is a zero-mean Gaussian random variable with covariance $\langle e_k, \mathcal{Q} e_k \rangle_G$. $\mathcal{P}_{j_n} \xi$ is a proper Gaussian element in the finite-dimensional space W_{j_n}, because

each $\xi(e_k)$ is Gaussian. Furthermore, notice that for any $u, v \in \mathbb{H}$,

$$\begin{aligned}\mathbb{E}\langle \mathcal{P}_{j_n}\xi, u\rangle\langle \mathcal{P}_{j_n}\xi, v\rangle =& \mathbb{E}\left[\left(\sum_{k\le j_n}\xi(e_k)\langle u, e_k\rangle\right)\left(\sum_{l\le j_n}\xi(e_l)\langle v, e_l\rangle\right)\right]\\ =& \sum_{k,l\le j_n}\langle u, e_k\rangle\langle v, e_l\rangle\langle e_k, \mathcal{Q}e_l\rangle_G\\ =&\langle \mathcal{P}_n u, \mathcal{Q}\mathcal{P}_n v\rangle_G.\end{aligned}$$

Since $\mathcal{P}_n = \mathcal{P}_n^*$, $\mathcal{P}_{j_n}\xi$ is a centred Gaussian with covariance $\mathcal{P}_{j_n}\mathcal{Q}\mathcal{P}_{j_n}^*$. Thus, the following expression makes sense,

$$\mathcal{P}_{j_n}X^{(n)} = \mathcal{P}_{j_n}\mathcal{A}\theta + \frac{1}{\sqrt{n}}\mathcal{P}_{j_n}\xi.$$

Besides, because of the first property in Assumption 4.9, we have

$$f^{(n)} = \mathcal{R}_n\mathcal{A}f = \mathcal{R}_n\circ\mathcal{P}_{j_n}\mathcal{A}f, \quad \forall f\in\Theta.$$

Now we claim that

$$\mathcal{R}_n\circ\mathcal{P}_{j_n}X^{(n)} = \theta^{(n)} + \frac{1}{\sqrt{n}}\mathcal{R}_n\circ\mathcal{P}_{j_n}\xi \tag{4.14}$$

is a well-defined Gaussian random element in Θ. It suffices to show the noise $\mathcal{R}_n\circ\mathcal{P}_{j_n}\xi$ is a proper random element in Θ. Denote $\mathcal{R}_n\circ\mathcal{P}_{j_n}$ by $\widehat{\mathcal{R}}_n$. Since

$$\begin{aligned}\mathbb{E}\|\widehat{\mathcal{R}}_n\xi\|_\Theta^2 =& \operatorname{Trace}(\mathcal{R}_n\mathcal{P}_{j_n}\mathcal{Q}\mathcal{P}_{j_n}\mathcal{R}_n^*) = \operatorname{Trace}\Big[(\mathcal{R}_n\mathcal{P}_{j_n}\mathcal{Q}^{1/2})(\mathcal{R}_n\mathcal{P}_{j_n}\mathcal{Q}^{1/2})^*\Big]\\ =&\|\mathcal{R}_n\mathcal{P}_{j_n}\mathcal{Q}^{1/2}\|_{HS}^2 = \sum_{k\le j_n}\|\mathcal{R}_n\mathcal{P}_{j_n}\mathcal{Q}^{1/2}e_k\|^2 \le \|\mathcal{R}_n\|^2\sum_{k\le j_n}\|\mathcal{P}_{j_n}(\mathcal{Q}^{1/2}e_k)\|^2\\ \lesssim& j_n\rho_n^2,\end{aligned}$$

where we use the properties of Schatten norms (see Proposition A.19 and Proposition A.20), and the fact that $\|\mathcal{P}_{j_n}(\mathcal{Q}^{1/2}e_k)\| \le \|\mathcal{Q}^{1/2}e_k\|_G \le \|\mathcal{Q}^{1/2}\|$. The preceding inequality shows that $\widehat{\mathcal{R}}_n X^{(n)}$ is indeed a proper random element in Θ. In addition, the weak second moment of the variable $\widehat{\mathcal{R}}_n\xi$ is

$$\begin{aligned}\sup_{\|\theta\|_0\le 1}\mathbb{E}\langle\theta, \widehat{\mathcal{R}}_n\xi\rangle_\Theta^2 =& \sup_{\|\theta\|_0\le 1}\langle\theta, \mathcal{R}_n\mathcal{P}_{j_n}\mathcal{Q}\mathcal{P}_{j_n}\mathcal{R}_n^*\theta\rangle_\Theta\\ =& \sup_{\|\theta\|_0\le 1}\|\mathcal{Q}^{1/2}\mathcal{P}_{j_n}\mathcal{R}_n^*\theta\|_G^2 = \|\mathcal{Q}^{1/2}\mathcal{P}_{j_n}\mathcal{R}_n^*\|_{\Theta;G}^2\\ =&\|\mathcal{R}_n\mathcal{P}_{j_n}\mathcal{Q}^{1/2}\|_{G;\Theta}^2 = \|\mathcal{R}_n\|_{G;\Theta}^2\|\mathcal{P}_{j_n}\mathcal{Q}^{1/2}\|^2 \simeq \rho_n^2,\end{aligned}$$

where the last inequality follows from the boundedness of $\mathcal{P}_{j_n}$ and $\mathcal{Q}^{1/2}$, and Assumption 4.9.

The first inequality shows that the first moment $\mathbb{E}\|\widehat{\mathcal{R}}_n\xi\|_0$ of the variable $\|\widehat{\mathcal{R}}_n\xi\|_0$ is bounded above by $\sqrt{j_n}\rho_n$. By Borell's inequality (see Lemma 3.11),

applied to the Gaussian random variable $\widehat{\mathcal{R}}_n\xi$ in Θ_0, we see that there exist positive constants a and b such that, for every $t > 0$,

$$\mathbb{P}\Big(\|\widehat{\mathcal{R}}_n\xi\|_0 > t + a\sqrt{j_n}\rho_n\Big) \leq e^{-bt^2/\rho_n^2}.$$

With $t = 2\sqrt{n}\eta_n/\sqrt{b}$, for η_n and ε_n satisfying (4.9) and (4.10) this yields, for some $a_1 > 0$,

$$\mathbb{P}\Big(\|\widehat{\mathcal{R}}_n\xi\|_0 > a_1\sqrt{n}\eta_n\Big) \leq e^{-4n\varepsilon_n^2}. \tag{4.15}$$

We apply this to bound the error probabilities of the tests

$$\tau_n = 1\{\|\widehat{\mathcal{R}}_n X^{(n)} - \theta_0\|_0 \geq M_0\eta_n\}, \tag{4.16}$$

where M_0 is a given constant, to be determined.

Under θ_0, the decomposition (4.14) is valid with $\theta = \theta_0$, and hence $\widehat{\mathcal{R}}_n X^{(n)} - \theta_0 = n^{-1/2}\widehat{\mathcal{R}}_n\xi + \theta_0^{(n)} - \theta_0$. By the triangle inequality it follows that $\tau_n = 1$ implies that $n^{-1/2}\|\widehat{\mathcal{R}}_n\xi\|_0 \geq M_0\eta_n - \|\theta_0^{(n)} - \theta_0\|_0$. By (4.7), the assumption that $f_0 \in H_\beta$ implies that $\|\theta_0^{(n)} - \theta_0\|_0 \leq M_1\delta(j_n, \beta)$, for some M_1, which is further bounded by $M_1\eta_n$, by assumption (4.10). Hence the probability of an error of the first kind satisfies

$$\mathbb{P}_{\theta_0}^{(n)}\tau_n \leq \mathbb{P}\Big(\frac{1}{\sqrt{n}}\|\widehat{\mathcal{R}}_n\xi\|_0 \geq (M_0 - M_1)\eta_n\Big).$$

For $M_0 - M_1 > a_1$, the right side is bounded by $e^{-4n\varepsilon_n^2}$, by (4.15).

Under θ the decomposition (4.14) gives that $\widehat{\mathcal{R}}_n X^{(n)} - \theta_0 = n^{-1/2}\widehat{\mathcal{R}}_n\xi + \theta^{(n)} - \theta_0$. By the triangle inequality $\tau_n = 0$ implies that $n^{-1/2}\|\widehat{\mathcal{R}}_n\xi\|_0 \geq \|\theta^{(n)} - \theta_0\|_0 - M_0\eta_n$. For θ such that $\|\theta - \theta_0\|_0 > M\eta_n$ and $\|\theta - \theta^{(n)}\|_0 \leq \eta_n$, we have $\|\theta^{(n)} - \theta_0\|_0 \geq (M-1)\eta_n$. Hence the probability of an error of the second kind satisfies

$$\mathbb{P}_\theta^{(n)}(1 - \tau_n) \leq \mathbb{P}\Big(\frac{1}{\sqrt{n}}\|\widehat{\mathcal{R}}_n\xi\|_0 \geq (M - 1 - M_0)\eta_n\Big),$$

For $M - 1 - M_0 > a_1$, this is bounded by $e^{-4n\varepsilon_n^2}$, by (4.15).

We can first choose M_0 large enough so that $M_0 - M_1 > a_1$, and next M large enough so that $M - 1 - M_0 > a_1$, to finish the proof. □