MASTER'S THESIS

Gaussian Process Approach to Behavioral Prediction

Tom Rosenström

University of Helsinki
Department of Mathematics and Statistics
Supervised by: Dos. Dario Gasbarra
December 2010

University of Helsinki

Department of Mathematics and Statistics

Rosenström, Tom: Gaussian Process Approach to Behavioral Prediction

Master's Thesis, 71 p.

Mathematics

December 2010

Abstract

A nonlinear regression problem of the type typically encountered in behavioral sciences, is considered. Here, one wishes to make a nonlinear prediction of real-valued outcome, using latent/unobserved predictor (independent, or explanatory) variable. Instead of actual predictor, one observes either a noisy version of it, or several indicators for it. These indicators could, for example, correspond to questionnaire items in standard psychometric measurement scale. Furthermore, it is assumed that little is known about the functional form of nonlinear regression function between the outcome and set of predictors.

The problem is approached by combining a standard latent variable model, Factor analysis, with regression based on Gaussian processes, and with new developments allowing "error-in-variables", or alternatively "noisy" predictors. Gaussian process regression takes statistical inference directly to the infinite-dimensional space of functions, and has gained increasing attention lately due to growing availability of computational capacities. This approach has been found effective when predictors are properly observed, with little measurement error. However, it turns out to be difficult to extend to the case of imprecisely observed predictors. Measurement error in variables is a rule rather than exception in behavioral sciences. In the current work, this problem is solved with an approach tailored to the needs of behavioral sciences. Presented solution can be used in corresponding situations in other fields of science.

Helsingin yliopisto

Matematiikan ja tilastotieteen laitos

Rosenström, Tom: Behavioraalinen ennustaminen Gaussisilla prosesseilla

Pro Gradu-tutkielma, 69 p.

Matematiikka, Stokastiikan linja

Joulukuu 2010

Abstrakti (In Finnish)

Tutkimme, mm. käyttäytymistieteissä esiintyvää, epälineaarista regressio-ongelmaa. Tavoitteena on tehdä epälineaarinen ennuste reaaliarvoiselle vastemuuttujalle käyttäen piileviä/ei-havaittuja ennustemuuttujia (”riippumattomia”/”selittäviä” muuttujia). Todellisen ennustemuuttujan sijaan, havaitaan vain kohinainen indikaattori, tai useita indikaattoreita. Nämä indikaattorit saattaisivat vastata esimerkiksi kyselylomakkeen kysymyksiä tunnetussa psykometrisessä mittarissa. Lisäksi oletetaan että vastemuuttujan ja ennustemuuttujien epälineaarisesta yhteydestä tiedetään hyvin vähän.

Yllä kuvattua ongelmaa lähestytään yhdistämällä standardi piilomuuttujamalli, Faktorianalyysi, Gaussisiin prosesseihin pohjaavaan regressiomenetelmään, sekä tällä alueella hiljattain saavutettuihin kohinaisten muuttujien mallintamisen edistysaskeliin. Regressio Gaussisilla prosesseilla vie tilastollisen päättelyn suoraan ääretönulotteiseen funktioavaruuteen, ja on viimeaikoina, kasvavan laskannallisen kapasiteetin myötä, herättänyt lisääntyvissä määrin kiinnostusta. Tämä lähestyminen on osoittautunut tehokkaaksi, kun ennustemuuttujat havaitaan virheettömästi, tai lähes virheettömästi. Se on kuitenkin osoittautunut hankalaksi laajentaa epätarkasti havaittujen ennustemuuttujien tapaukseen. Mittausvirhe muuttujissa on ennemmin sääntö kuin poikkeus käyttäytymistieteissä. Tässä työssä epätarkkojen havaintojen ongelma ratkaistaan käyttäytymistieteiden tarpeisiin räätälöidyllä lähestymisellä. Esitetty ratkaisu on sovellettavissa muiden alojen vastaaviin tilanteisiin.

Contents

Preface and Acknowledgements

Some years ago I ran into the concept of Gaussian process regression. Probably it was via the free web-books of David MacKay, and Carl Edward Rasmussen and Christopher Williams. For some reason, the idea of doing statistical inference directly in infinite-dimensional function spaces stroke me as fascinating. At that time, I had little reasons to dig deeper into the matter. Once I started my present work at the Faculty of behavioral sciences, which involves loads of regression analysis, this approach immediately came to mind. I was working with optimal behavioral regression prediction of an atherosclerosis indicator and, at that time, I could not see what, in this context, could be more optimal than Gaussian process regression. Very rapidly I discovered that it does not perceivably outperform linear methods, even though we (including my collaborators) strongly suspected that nonlinearities should lie hidden in these data. After some digging, I realized that measurement error in predictor variables inflicts a lot larger effect on non-linear model estimation than for linear. I believe this is primary reason for the failure of Gaussian process regression in this case. This master's thesis offered a good reason to do a theoretical investigation to the question "can Gaussian process methods extend to this case as well"? Hence, it answers to the dual need of graduation and solving a practical problem. Whether the presented solution is a good one, or not, is still a matter of empirical work. In any case, I learned a lot of practical mathematics while doing this thesis.

I thank my supervisor Dario Gasbarra for providing help, conversation and useful materials. I also thank professor Liisa Keltikangas-Järvinen for encouraging open-minded pursue of new methods, and Miika Pihlaja for constructive comments and typo hunting.

1 Introduction

1.1 The behavioral setting

Personality is a time-average of behavior. This construct has shown success in predicting future behavior, and clear associations with health outcomes, such as atherosclerosis and depression (John, Robins & Pervin 2008; Hintsanen, et al., 2009; Puttonen et al., 2008). This thesis is an attempt for a mathematical solution to some central challenges that arise when one tries to predict some outcome from current behavioral measures (e.g. Cloninger, Przybeck, Svrakic & Wetzel, 1993; Costa & McCrae, 1985).

Personality, or typical/average behavior of an individual, is often described as normally distributed deviations from the population average of behavior. It is measured by asking several questions that are thought to reflect between individual variation in some latent (hidden) trait. A latent trait could, for example, describe how enthusiastic one is about various social interactions. Depending on the theory, there are 3 to 7 separate traits, or dimensions, where individuals gain values approximately according to a normal distribution. Any given trait alone is an incomplete description of the behavioral profile of an individual, and may manifest in different behavioral outcomes according to values of the other traits/dimensions. Thus, effects or outcomes, depend on interactions between these dimensions (real-valued variables in practice). In addition, dependence is unlikely to be monotonic. Instead, it is typical that both extremes of any trait are somehow problematic, because individual with extreme value, high or low, behaves very differently from the general population, with respect to this trait. Deviant behavior may result in adjustment difficulties, which in turn may lead to psychosocial stress and pernicious habits.

To collect the methodological implications so far, modeling of behavioral outcomes with personality measures requires a nonlinear regression model with several, highly interacting, independent variables, or dimensions. However, one rarely has very good intuition to functional shapes of these nonlinearities and interactions. Hence, it would be advantageous to be able to learn these from the data. It is plausible to assume that outcomes, like depression or atherosclerosis in population, depend on these traits in a continuous manner, if at all. Thus, an assumption of continuity could be used to smooth out the dependence of outcome Y on behavioral traits $X = (X_1, \ldots, X_d)$, that is to estimate the conditional expectation of Y given X, $E[Y|X]$. Here, and whenever we wish to emphasize this distinction, capital letters are used to denote random variables,

and corresponding lower-case letters denote their observed values. $\mathbb{R}^d := \mathbb{R} \times \mathbb{R} \times \ldots \times \mathbb{R}$ (d-copies) is a standard d-dimensional Euclidian space. $A := B$ is used to define A as B.

Nonlinear regression, with a minimal assumption of continuity, can be based on Gaussian stochastic processes (Rasmussen & Williams, 2006). Third chapter of this thesis will describe how this is done, and why it might be an efficient approach. Second chapter is devoted to introducing Gaussian stochastic processes, or alternatively Gaussian fields. Intuitive idea behind this approach is to set up a random field indexed by personality dimensions, condition it on the observations, and then study how value of the field (distribution of outcome variable) varies in different parts of the indexing space (behavioral traits, independent variables). As the measured values of the behavioral traits for two individuals represent the multidimensional distance between their behavior, it is natural to think of outcome as a random field indexed by the traits (getting values in $\mathbb{R}^d$). Thus, the outcome is modeled as a variable whose mean indicates the typical value for those with a given profile/configuration of behavioral traits, and whose variance corresponds to between individual variation and measurement error in outcome.

While such approach have been previously used to solve problems in the fields of Machine learning and Geology, naïve approach to Gaussian process regression (GPR) does not work well for behavioral assessment. Observations from the indexing variable cannot be obtained without significant amounts of observation error, also referred as measurement error. It will be shown in the chapter three, that this brakes down the standard approach. Error-in-variables setting has been listed as open challenges of GPR-research (Rasmussen & Williams, 2006). Recently, an attempt for work-around was suggested for this problem (Girard, 2004; Dallaire, Besse & Chaib-draa, 2009), and we will also utilize it in the chapter five. However, this solution can be enhanced with a more direct estimate for the measurement error variance. Such an estimate will be first derived in chapter four, using a psychometric measurement framework (e.g. Tarkkonen & Vehkalahti, 2005). Chapters five and six collect all the above pieces and constructs a data-driven (assumption deprived) non-linear regression/smoothing method for the behavioral outcomes.

1.2 Mathematical background

Necessary background material for this work consists of the measure theoretic probability and linear algebra, as taught at the University level. Some familiarity with statistics is helpful also. To activate memory of the reader, and to review some notation, we briefly cover few standard concepts. A collection, $(\Omega, \mathcal{F}, P)$, of the set Ω, its σ-algebra $\mathcal{F}$, and a probability measure P on $\mathcal{F}$, is called a *probability space*. Sometimes an equivalent term "σ-field" is used, instead of σ-algebra. Both refer to a collection of sets that includes an empty set and is closed under complements and countable unions of its members. σ-

algebra is *generated* by the collection of sets τ, when it is the smallest collection of sets that, both, is a σ-algebra and contains all sets in τ. When τ is a *topology*, that is collection all open sets of the space, σ-algebra generated by τ is called the *Borel σ-algebra*, and denoted as $\mathcal{B}$.

A function $X : \Omega \to \mathbb{R}^d$ is *$\mathcal{F}$-measurable* if for all Borel sets $B \in \mathcal{B}$ (or equivalently, for all open sets B)

$$X^{-1}(B) := \{\omega \in \Omega; X(\omega) \in B\} \in \mathcal{F}.$$

$\mathcal{F}$-measurable function is a *random variable*. Here we stick with real-valued random variables, but range of X does not need to be real-valued space in general probability theory. σ-algebra generated by X is the smallest σ-algebra on Ω that contains all sets $\{X^{-1}(B); B \in \mathcal{B}\}$. We say that any map X (random variable or not) is a *measurable map* from *measurable space* $(\Omega, \mathcal{F})$ to another measurable space $(\Omega', \mathcal{F}')$ if $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{F}'$. Here, Ω and Ω' are sets, and $\mathcal{F}$ and $\mathcal{F}'$ are σ-algebras on them. Often we leave measurable spaces undefined when they are clear from the context, or statement generalizes to any space, and just speak of measurable maps. For real-valued sets, associated σ-algebra is always taken to be their Borel σ-algebra. Compositions, sums, subtractions, products, divisions, infimum, supremum and limit infimum/supremum of measurable maps are all again measurable maps (e.g. Klenke, 2008, Chap. 1). Notice that we have not defined Ω and it may be almost any set. Here, we will soon face a situation where Ω is a set of functions from $T \subset \mathbb{R}^n$ to one-dimensional real-space $\mathbb{R}$.

Every random variable X *induces* a probability measure on the set where it gets values. If $X : \Omega \to \mathbb{R}^d$, then the measure

$$\mu_X(B) := P_X(B) := (P \circ X^{-1})(B) = P(X^{-1}(B)), \qquad B \in \mathcal{B} \tag{1.1}$$

defines the *distribution* of X. Expectation $E[X]$ of X is the integral of X with respect to measure μ_X, that is

$$E[X] := \int_\Omega X(\omega) dP(\omega) = \int_{\mathbb{R}^d} x d\mu_X(x). \tag{1.2}$$

Indicator function, $\omega \mapsto 1_B(\omega) \in \{0, 1\}$, gains a value 1 when $\omega \in B$, and 0 otherwise. Integral of X over the set B can be shortly expressed as $E[1_B X]$ (characteristic function, χ_B, plays the same role in Analysis, but refers to different function in Probability theory). Note that it is standard practice to omit the argument ω of random variable X from the notations, while it is always implicitly present when speaking of random variables, or vectors. Above concepts, as well as the concepts like conditional expectation, should be familiar to the reader. The rest of the thesis builds on them.

The conditional expectation can be defined via Radon-Nikodym theorem. A measure ν is said to be absolutely continuous with respect to measure μ, if for all $A \in \mathcal{F}$, $\nu[A] = 0$ always when $\mu[A] = 0$. This is denoted as $\nu \ll \mu$. One form of the Radon-Nikodym theorem states the following

Theorem 1.1 (Radon-Nikodym theorem)**.** *If, and only if,* $\nu \ll \mu$*, then*

$$\nu[A] = \int_A Z(\omega) d\mu[\omega], \qquad A \in \mathcal{F},$$

for some almost surely unique measurable map $Z : \Omega \to \mathbb{R} \cup \{\infty\}$*.*

Proof. See e.g. corollary 7.34 in (Klenke, 2008).

□

Now, if $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{F}$, for any positive random variable X, we can set $\nu[A] := E[X1_A]$ and $\mu[A] := P[A|\mathcal{G}] = P[A]$ for all $A \in \mathcal{G}$. Then, $E[X1_A] = E[Z1_A]$, where Z is $\mathcal{G}$-measurable random variable. From this, Kolmogorov made a natural definition: $Z =: E[X|\mathcal{G}]$. To see that this is natural, notice that $E[X|\mathcal{G}]$ is random until we have observed some event $A \in \mathcal{G}$. When A has happened, we want the conditional expectation to refer to expectation of events within the set A, that is, expectation given that A is already known. Extension of the definition to general random variables, instead of positive ones, is straightforward.

Although, previous knowledge of L^p-spaces is not necessary, we sometimes refer to them. When $p \geq 1$ and $E[|X|^p] < \infty$, random variable X is said to belong to L^p-space. The size of this space depends on Ω, $\mathcal{F}$ and P, and it is often referred as $L^p(\Omega, \mathcal{F}, P)$. L^p-spaces are Banach-spaces, and L^2 is an inner product (Hilbert) space, with inner product $\langle X, Y \rangle := E[XY]$ (Klenke, 2008, chap. 7). If $X_n \to X$ as $n \to \infty$, and $X_n \in L^p(\Omega, \mathcal{F}, P)$ for all n, then also $X \in L^p(\Omega, \mathcal{F}, P)$. Here, the convergence is with respect to L^p-norm $\| \cdot \|_p$ defined as $\| X \|_p := E[|X|^p]^{1/p}$. This is a general property of Banach-spaces.

2 Gaussian random fields

2.1 Stochastic process

We begin by defining a stochastic process.

Definition 2.1 (Stochastic process)**.** *Stochastic process is a parameterized collection of random variables* $\{Y_t : \Omega \to \mathbb{R}^k; t \in T\} =: \{Y_t\}$ *defined on a probability space* $(\Omega, \mathcal{F}, P)$ *and assuming values in* $\mathbb{R}^k$*.*

T could be fairly general manifold (see e.g. Adler & Taylor, 2007), but for us it suffices that $T \subseteq \mathbb{R}^d$, for some $d \in \mathbb{N}$. Most typical interpretation is that T represents (1-dimensional) time. We mostly consider $\mathbb{R}$-valued ($k = 1$) processes. While for each fixed $t \in T$, Y_t provides a random variable $\omega \mapsto Y_t(\omega)$, for each fixed $\omega \in \Omega$, it provides a *function* $t \mapsto Y_t(\omega)$. Because, elements of Ω are functions, it can be thought of as a function space. Thus, the elemental events are functions from T to $\mathbb{R}$, and more general events in event-space Ω are sets of functions. Here the index set $T \subset \mathbb{R}^d$ will take the role of predictor, or input, variables (behavioral traits), while values of $\{Y_t\}$ will represent the outcome of interest. Typically, when T has more than one dimension, $\{Y_t\}$ is also called a *random field* (and sheet, when $d = 2$). Here, we anticipate our atypical indexing variable, by switching to x, a common symbol for predictor in the regression context. A stochastic process $\{Y_x\}$ is said to be *centered* if $E[Y_x] = 0$ for all x. For reasons that become obvious later, we can concentrate on the centered fields.

$\{Y_x\}$ readily defines a distribution for a set of n observations,

$$\{(Y_{X_1}, X_1), \ldots, (Y_{X_n}, X_n)\},$$

but given a set of observations and a probability measure on them, can we always construct a stochastic process that yields this finite-dimensional distribution? This knowledge is central for the modeling approach chosen here, and it is provided by the Kolmogorov's extension theorem, given two natural consistency requirements.

Theorem 2.1 (Kolmogorov's extension theorem)**.** *If* $\mu_{x_1,\ldots,x_n}$ *is a probability measure on* $\mathbb{R}^n$ *(or* $\mathbb{R}^{nk}$ *if* Y_t *is* $\mathbb{R}^k$*-valued),* $\{B_1 \in \mathcal{B}_1, \ldots, B_n \in \mathcal{B}_n\}$ *and*

$$\mu_{x_{\sigma(1)},\ldots,x_{\sigma(n)}}(B_1 \times \cdots \times B_n) = \mu_{x_1,\ldots,x_n}(B_{\sigma^{-1}(1)} \times \cdots \times B_{\sigma^{-1}(n)})$$

for all permutations σ *on* $\{1, \ldots, n\}$ *and*

$$\mu_{x_1,\ldots,x_n}(B_1 \times \cdots \times B_n) = \mu_{x_1,\ldots,x_n,x_{n+1},\ldots,x_{n+m}}(B_1 \times \cdots \times B_n \times \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{m \text{ copies}}),$$

for all $m \in \mathbb{N}$*, then there exist a probability space* $(\Omega, \mathcal{F}, P)$ *and a stochastic process* Y_x *such that*

$$\mu_{x_1,\ldots,x_n}(B_1 \times \cdots \times B_n) = P(Y_{x_1} \in B_1, \cdots, Y_{x_n} \in B_n)$$

for all $\{x_1, \ldots, x_n\}, n \in \mathbb{N}$*, and for all borel sets* $B_1, \ldots, B_n$*.*

Proof. e.g. Theorem 14.36 in Klenke, (2008).

□

2.2 Gaussian process

From here onwards, $(\cdot)^T$ denotes a transpose of a matrix or vector. For matrix-valued input, $|\cdot|$, denotes the absolute value of determinant. A real-valued random vector $X : \Omega \to \mathbb{R}^d$ is Gaussian, or normally distributed, when it has Lebesgue-integrable probability density function of the form

$$\varphi_X(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}, \tag{2.1}$$

or a characteristic function

$$\theta \mapsto E[e^{iX^T\theta}] = e^{i\mu^T\theta - \frac{1}{2}\theta^T\Sigma\theta}, \qquad i = \sqrt{-1}. \tag{2.2}$$

Here Σ is the covariance matrix with elements

$$Cov[X_k, X_j] := E[(X_k - E[X_k])(X_j - E[X_j])], \qquad k, j \in \{1, \ldots, d\},$$

and $\mu \in \mathbb{R}^d$ is the mean vector $E[X]$. Strictly speaking, only the latter condition is required, while density function may not always exist. This distribution is denoted as $N_d(\mu, \Sigma)$, and $X \sim N_d(\mu, \Sigma)$ means that X is distributed accordingly. Subscript d may be dropped if the dimension is self-evident from the context.

Gaussian variables play a fundamental role in probability theory, and they have several desirable analytical properties. While this distribution has deep connections to asymptotics (large samples), entropy, etc., its convenient analytical properties are central to the current work. They are the main reason why we want to analyze Gaussian processes, a subset of stochastic processes.

Definition 2.2 (Gaussian process)**.** *An* $\mathbb{R}^k$*-valued stochastic process* $\{Y_x\}$ *is Gaussian if all of its finite-dimensional samples/projections* $(Y_{x_1}, Y_{x_2}, \ldots, Y_{x_n}), n \in \mathbb{N}$ *are.*

In above definition, $(Y_{x_1}, Y_{x_2}, \ldots, Y_{x_n})$ is $\mathbb{R}^{nk}$-dimensional normally distributed random variable. Dimension of $x \mapsto Y_x$ is infinite, as is the case for functions in general, but Gaussian process is defined trough its projections/samples. As the characteristic function completely determines a distribution of the random variable, from equation 2.2 one sees that any finite-dimensional Gaussian distribution is completely determined by its mean and covariance matrix. Zero covariance implies independence for the jointly normally distributed random variables. As for any sample of a real-valued process, the (i, j)-th element of the covariance matrix is given by $Cov[Y_{x_i}, Y_{x_j}]$, and the mean vector by $(E[Y_{x_1}], \ldots, E[Y_{x_n}])$, the *entire process* is defined by the mean and covariance functions,

$$x \mapsto E[Y_x] =: f(x)$$

and

$$(x, x') \mapsto Cov[Y_x, Y_{x'}] =: C(x, x').$$

This is truly a nice analytic property.

Most studied Gaussian process is, without doubt, Brownian motion. Brownian motion can, for example, be used to model the motion of small particles in the fluid as a function of time t. The covariance function of this process is $C(t, t') = \min(t, t')$. There exists a version of Brownian motion that almost surely (for almost all ω, outside of sets of measure zero) has continuous paths, that is, elements $\omega \in \Omega$ are continuous functions. For us, continuity is desirable property. However, paths of Brownian motion are so "rugged", or irregular, as to be (almost surely) nowhere differentiable (Klenke, 2008, chap. 21). Figure 3.1 shows one approximate path of a Brownian motion. Actually, any function C, with $\sum_{ij} v_i v_j C(x_i, x_j) \geq 0$ for all finite sets of points $\{x_i\}$ and arbitrary real coefficients v_i, is a covariance function of some Gaussian process. These functions are called non-negative definite, because for any finite sample $(x_1, x_2, \ldots, x_n)$ they yield positive non-negative definite covariance matrices $C := [C(x_i, x_j)]_{i,j=1}^n$. A matrix C is called non-negative definite, if $v^T C v \geq 0$ for all $v \neq 0 \in \mathbb{R}^n$, and positive definite if $\geq$ is replaced with strict inequality.

Theorem 2.2 (Covariance functions)**.** *Let C be any function $C : \chi \times \chi \to \mathbb{R}$ in any index-space χ. If C satisfies the inequality*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} v_i v_j C(x_i, x_j) = v^T C v \geq 0 \tag{2.3}$$

for any sample $(x_1, \ldots, x_n) \in \chi^n$, and any vector $(v_1, \ldots, v_n)^T =: v \in \mathbb{R}^n$ such that $v \neq 0$, then C is a covariance function of some $\mathbb{R}$-valued Gaussian process on χ.

Proof. Consider first the situation: $v^T C v > 0$ for all $v \neq 0$. For the finite sample, $(x_1, \ldots, x_n)$, mean 0 and covariance C define an $\mathbb{R}^n$-dimensional centered Gaussian distribution/measure. Then according to Kolmogorov's extension theorem, there is a corresponding stochastic process. Its covariance

function is $C(\cdot,\cdot)$, and hence for any other finite sample $v^TCv > 0$ holds, and the Gaussian measure exists.

If $v^TCv = 0$ for some $v \neq 0$, then there is no variance in the direction spanned by v, and no density function 2.1, but there still exists a Gaussian distribution. This follows from the fact that the set of Gaussian distributions is closed in $L^2(\Omega)$. It contains all converging limits, or equivalently, all points in its closure. This follows from the continuity of the Gaussian characteristic function. If (μ_k, Σ_k) are sequence of means and covariances of a corresponding sequence of Gaussian variables (Y_k), and $(\mu, \Sigma) := (\lim_{k\to\infty} \mu_k, \lim_{k\to\infty} \Sigma_k)$, then by the continuity

$$\lim_{k\to\infty} e^{i\mu_k^T\theta - \frac{1}{2}\theta^T\Sigma_k\theta} = e^{i\mu^T\theta - \frac{1}{2}\theta^T\Sigma\theta},$$

which is again Gaussian. To see that there is a positive definite sequence converging to C for which $v^TCv = 0$, but $v \neq 0$, take $C_k := C + Ik^{-1}$, where I is an identity matrix, and $Y_k \sim N(0, C_k)$. Then if $v \neq 0$, $v^TIk^{-1}v > 0$, for any integer k, and $\lim_{k\to\infty} v^TC_kv = v^TCv = 0$.

□

To gain some intuition for what a degenerate Gaussian distribution without density might be, consider one-dimensional case with distribution $N_1(0, \frac{1}{k})$ as k tends to infinity. For every larger k, Y_k is more and more certainly in the vicinity of 0. After the limiting process, only possible distribution function is the Heaviside step-function H for which $H((-\infty, x]) = 1$ when $x \geq 0$ and $H((-\infty, x]) = 0$ otherwise. But, according to Radon-Nikodym theorem, for there to be a Lebesgue-integrable density φ for which $\int_{-\infty}^{x} \varphi(x)dx = H((-\infty, x])$ and $\int_{\mathbb{R}} \varphi(x)dx = 1$, where dx is the Lebesgue-measure, it should hold that $\int_A dx = 0$ implies $H(A) = 0$, for any set $A \in \mathcal{B}$. This is not true for the set $A = \{0\}$, whose Lebesgue-measure is 0, but

$$H(A) = H((-\infty, 0]) - \lim_{t\uparrow 0} H((-\infty, t]) = 1.$$

3 Gaussian process regression (GPR)

3.1 Regression

Most elementary form of regression model, is the Linear regression model. In this model, the outcome variable Y is thought as a linear sum of appropriately weighted predictor variables (traits) plus some error variation, ξ. That is, $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d + \xi := f(X) + \xi$, where $f : \mathbb{R}^d \to \mathbb{R}$, $f(X) = \sum_{i=0}^{d} \beta_i X_i$, with $X_0 := 1$. As already visible, the task of regression generalizes to task of finding some function f which describes the systematic (X-dependent) part of outcome as a function of input, or predictor, variables.

Definition 3.1 (Regression)**.** *Regression of outcome Y to input X is a process of finding such function f that for any other function g, within some class of functions, $\mathcal{L}(Y, f(X)) \leq \mathcal{L}(Y, g(X))$ holds, where $\mathcal{L}$ is some loss function ($\mathcal{L} \geq 0$).*

The loss function $\mathcal{L}$ could get many forms, but most used tends to be the squared loss $\mathcal{L}(Y, f(X)) := E[(Y - f(X))^2]$, also known as the mean of squared error. We will also use squared loss in the following. With this loss function, it can be shown that optimal choice for the regression function f is the conditional expectation of Y given X, $E[Y|X] := E[Y|\sigma(X)]$, where $\sigma(X)$ denotes the σ-algebra generated by X. Let 1_B be the indicator function of the set B. In general, conditional expectation of Y given information (σ-algebra) $\mathcal{F}$ is the (almost surely) unique $\mathcal{F}$-measurable random variable $E[Y|\mathcal{F}]$ for which $E[1_B E[Y|\mathcal{F}]] = E[1_B Y]$ for all $B \in \mathcal{F}$. If Y already is $\mathcal{F}$-measurable, $E[Y|\mathcal{F}] = Y$ and $E[YX|\mathcal{F}] = YE[X|\mathcal{F}]$. Furthermore, the "Tower property" (Klenke, 2008), or equivalently, "Law of iterated expectation", holds. Although, familiarity with conditional expectation was assumed, this result is central in the following, and we will prove it. In fact, a more general result, from which the Tower property follows, is proved.

Theorem 3.1. *If $\mathcal{F}$ and $\mathcal{G}$ are two σ-algebras, and Y a random vector, then $E[E[Y|\mathcal{F}]|\mathcal{G}] = E[Y|\mathcal{F} \cap \mathcal{G}]$.*

Proof. Intersection (but not usually union) of two σ-algebras is again a σ-algebra. Left-hand side of the equation is both, $\mathcal{F}$- and $\mathcal{G}$-measurable, that is, $\mathcal{F} \cap \mathcal{G}$-measurable. In other words, for $B \in \mathcal{B}$, if $E[E[Y|\mathcal{F}]|\mathcal{G}]^{-1}(B) \in \mathcal{F}$ and $E[E[Y|\mathcal{F}]|\mathcal{G}]^{-1}(B) \in \mathcal{G}$, then $E[E[Y|\mathcal{F}]|\mathcal{G}]^{-1}(B) \in \mathcal{F} \cap \mathcal{G}$. By the definition,

$E[Y|\mathcal{F}\cap\mathcal{G}]$ is that almost surely unique random vector which satisfies for every $B \in \mathcal{F} \cap \mathcal{G}$

$$E[1_B E[Y|\mathcal{F} \cap \mathcal{G}]] = E[1_B Y].$$

We must now show that $E[E[Y|\mathcal{F}]|\mathcal{G}]$ is this same unique random vector. Let B be any set in $\mathcal{F}\cap\mathcal{G}$. Then also $B \in \mathcal{F}$ and $B \in \mathcal{G}$. By applying the definition of the conditional expectation twice, it follows that

$$E[1_B E[E[Y|\mathcal{F}]|\mathcal{G}]] = E[1_B E[Y|\mathcal{F}]] = E[1_B Y].$$

□

The Tower property then follows from the fact that if $\mathcal{G} \subset \mathcal{F}$, then $\mathcal{F} \cap \mathcal{G} = \mathcal{G}$.

Corollary 3.1 (Tower property)**.** *If $\mathcal{F}$ and $\mathcal{G}$ are σ-algebras, and $\mathcal{G} \subset \mathcal{F}$, then $E[E[X|\mathcal{F}]|\mathcal{G}] = E[E[X|\mathcal{G}]|\mathcal{F}] = E[X|\mathcal{G}]$.*

Tower property shows that one may compute conditional expectations in iterated manner. It also shows that most "coarse" information (smaller collection of sets), or least random information (in probabilistic terms), is preserved when taking expectations.

Next result shows that the conditional expectation of Y given X is the optimal prediction of Y using X, in the sense of the mean of squared error. Consider observations of the form $y = f(x) + \xi$ that are sums of realizations of some measurable function of a random variable $X : \Omega_X \to \mathbb{R}^d$ and some error variable $\Xi : \Omega_\Xi \to \mathbb{R}$.

Theorem 3.2. *If $Y \in \mathbb{R}$ is $\sigma((X,\Xi))$-measurable and $X \in \mathbb{R}^d$ is $\sigma(X)$-measurable, with $E[Y^2], E[X^2] < \infty$, then $\sigma(X) \subset \sigma((X,\Xi))$ and*

$$E[(Y - f(X))^2] \geq E[(Y - E[Y|\sigma(X)])^2]$$

for all Lebesgue-measurable functions f.

Proof. If Ω_Ξ is the domain of Ξ and we identify all sets $B \in \sigma(X)$ with (B, Ω_Ξ), it is clear that we get the same σ-algebra as $\sigma(X)$. On the other hand, $(B, \Omega_\Xi) \in \sigma(X, \Xi)$ for all B, and hence, $\sigma(X) \subset \sigma(X, \Xi)$. Since f is measurable, composition $f \circ X = f(X)$ is $\sigma(X)$-measurable. From the Jensen's inequality, $E[E[Y|\sigma(X)]^2] \leq E[E[Y^2|\sigma(X)]] = E[Y^2]$, giving the integrability of the conditional expectation. Because $E[Y|\sigma(X)]$ and $f(X)$ are $\sigma(X)$-measurable, it follows from the properties of conditional expectation that $E[YE[Y|\sigma(X)]] = E[E[YE[Y|\sigma(X)]|\sigma(X)]] = E[E[Y|\sigma(X)]^2]$, and that $E[f(X)Y] = E[YE[f(X)|\sigma(X)]]$. Hence,

$$\begin{aligned}
&E[(Y - f(X))^2] - E[(Y - E[Y|\sigma(X)])^2] \\
&\quad = E[Y^2 - 2Yf(X) + f(X)^2 - Y^2 + 2YE[Y|\sigma(X)] - E[Y|\sigma(X)]^2] \\
&\quad = E[f(X)^2 - 2Yf(X) + E[Y|\sigma(X)]^2] \\
&\quad = E[(f(X) - E[Y|\sigma(X)])^2] \geq 0
\end{aligned}$$

□

Above theorem shows also that conditional expectation is an L^2-projection from space $L^2(\Omega_{(X,\Xi)}, \sigma(X,\Xi), P_{(X,\Xi)})$ to space $L^2(\Omega_X, \sigma(X), P_X)$, since it is known from the Functional analysis that there is a unique such projection, and it has smallest L^2-distance to any $Y \in L^2(\Omega_{(X,\Xi)}, \sigma(X,\Xi), P_{(X,\Xi)})$.

3.2 Regression with a Gaussian process

How does one then perform regression *with a Gaussian process*, that is, Gaussian process regression (GPR). For this we need below result about the conditionals of Gaussian distributions.

Theorem 3.3 (Conditional distributions of a Gaussian)**.** *Let an $\mathbb{R}^d$ -valued random vector $X = (X_1, \ldots, X_{d_1}, X_{d_1+1}, \ldots, X_{d_1+d_2})$ be distributed as $N_d(\mu, C)$, where $d = d_1 + d_2$, and $\mu = (\mu_1, \mu_2)$ with $\mu_1 \in \mathbb{R}^{d_1}$ and $\mu_2 \in \mathbb{R}^{d_2}$. Let $C_{11} \in \mathbb{R}^{d_1 \times d_1}$, $C_{22} \in \mathbb{R}^{d_2 \times d_2}$, $C_{12} \in \mathbb{R}^{d_1 \times d_2}$ and $C_{21} \in \mathbb{R}^{d_2 \times d_1}$ with*

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \in \mathbb{R}^{d \times d}.$$

Then, conditional vector $(X_1, \ldots, X_{d_1} | X_{d_1+1}, \ldots, X_d)$ is distributed as $N_{d_1}(\mu_{1|2}, C_{1|2})$, where

$$\mu_{1|2} = \mu_1 + C_{12}C_{22}^{-1}((X_{d_1+1}, \ldots, X_d)^T - \mu_2) \tag{3.1}$$

$$C_{1|2} = C_{11} - C_{12}C_{22}^{-1}C_{21} \tag{3.2}$$

Proof. The density function of multinormal (Gaussian) distribution is

$$\varphi_X(x) = \frac{1}{(2\pi)^{d/2}|C|^{1/2}} e^{-\frac{1}{2}(x-\mu)C^{-1}(x-\mu)^T},$$

where $|C|$ is the absolute value of determinant of C, $|\det(C)|$. Recall that, for the determinant following equalities hold: $|C^{-1}| = 1/|C|$, $|C| = |C^T|$, and $|C||A| = |CA|$ when multiplication is defined. Let A be the matrix of invertible linear transformation $A : \mathbb{R}^d \to \mathbb{R}^d$. Then, according to Change of variables formula of probability calculus, the density of $Y = AX$ is

$$\begin{aligned}
\varphi_Y(y) &= \varphi_X(A^{-1}y)|D_y(A^{-1}(y))| \\
&= \frac{1}{(2\pi)^{d/2}|C|^{1/2}} e^{-\frac{1}{2}(A^{-1}y-\mu)^T C^{-1}(A^{-1}y-\mu))}|A^{-1}| \\
&= \frac{1}{(2\pi)^{d/2}(|A||C||A^T|)^{1/2}} e^{-\frac{1}{2}(y-A\mu)^T A^{-T}C^{-1}A^{-1}(y-A\mu))} \\
&= \frac{1}{(2\pi)^{d/2}|ACA^T|^{1/2}} e^{-\frac{1}{2}(y-A\mu)^T (ACA^T)^{-1}(y-A\mu))}.
\end{aligned}$$

Thus, Y is distributed as $N_n(A\mu, ACA^T)$. Now, set $A = \begin{pmatrix} I_{d_1} & -C_{12}C_{22}^{-1} \\ 0 & I_{d_2} \end{pmatrix}$, where I_d denotes $d \times d$ identity matrix. Make the partitions

$$(X^{(1)}, X^{(2)}) = ((X_1, \ldots, X_{d_1}), (X_{d_1+1}, \ldots, X_d)),$$
$$Y = (Y^{(1)}, Y^{(2)}) =: (X^{(1)} - C_{12}C_{22}^{-1}X^{(2)}, X^{(2)}) = AX.$$

Then, Y is Normally distributed with covariance

$$ACA^T = \begin{pmatrix} C_{11} - C_{12}C_{22}^{-1}C_{21} & 0 \\ 0 & C^{22} \end{pmatrix},$$

implying that $Y^{(1)}$ and $Y^{(2)} = X^{(2)}$ are independent. Thus, we can write $\varphi_Y(y)dy = \varphi_{Y^{(1)}}(y^{(1)})dy^{(1)}\varphi_{Y^{(2)}}(y^{(2)})dy^{(2)}$ for the density of Y, giving $\varphi_{Y^{(1)}|Y^{(2)}}(y^{(1)}) = \varphi_{Y^{(1)}|X^{(2)}}(y^{(1)}) = \varphi_{Y^{(1)}}(y^{(1)})$. Finally, since $X^{(1)} = Y^{(1)} + C_{12}C_{22}^{-1}X^{(2)}$, we get the desired conditional density by translating above distribution with a constant $C_{12}C_{22}^{-1}x^{(2)}$

$$\varphi_{X^{(1)}|X^{(2)}=x^{(2)}}(x^{(1)}) = \varphi_{Y^{(1)}}(y^{(1)} + C_{12}C_{22}^{-1}x^{(2)}). \tag{3.3}$$

Now, $\varphi_{X^{(1)}|X^{(2)}=x^{(2)}}$ is still Gaussian, and its covariance is given by equation 3.2, and mean by $E[y^{(1)}] + C_{12}C_{22}^{-1}X^{(2)}$, which equals right side of equation 3.1.

□

With above result we can show how to perform regression with a Gaussian process (assuming positive definite covariance). If we have n real-valued outcomes $(Y_1, \ldots, Y_n)$, we may consider them as drawn from centered Gaussian field where the field position is indicated by corresponding input/predictor variables $(X_1, \ldots, X_n)$. As every finite-dimensional sample from this process is Gaussian, distribution of the outcome is entirely determined by the input data and covariance function of the Gaussian process. By choosing an appropriate covariance function for (Y_i, Y_j) we are essentially assuming a model for the data. However, this assumption is not very strict. We do not assume linearity, or other fixed functional form, but instead make an assumption regarding the smoothness properties of the data. For example, it is often reasonable modeling assumption that outcome values (Y_i, Y_j) are close to each other when predictor variables (X_i, X_j) are. For random variables, Y_i is "close" to Y_j when their covariance is high. Thus, we could fix the covariance function to be, for example, the squared exponential form, often used in the literature.

Definition 3.2 (Squared exponential covariance function)**.** *If (Y_i, Y_j) are centered real-valued outcome random variables and (x_i, x_j) are corresponding $\mathbb{R}^d$-dimensional predictor/input values, set*

$$C(x_i, x_j) := E[Y_iY_j] = ve^{-\frac{1}{2}(x_i - x_j)^T W^{-1}(x_i - x_j)}, \tag{3.4}$$

with positive constants $v \in \mathbb{R}_+$ and diagonal matrix $W \in (\mathbb{R}_+ \cup \{0\})^{d \times d}$. Then, this function is called the Squared exponential covariance function.

Here, i^{th} diagonal element of W provides the characteristic length-scale of component X_i. It describes how fastly Y can change as a function of X_i. v gives the overall variance of Y (set $C(x, x)$ to see this). We will use this covariance function as an example, but lots of other useful covariance functions exist for different situations (Rasmussen & Williams, 2006, chap. 4). Figure 3.1 demonstrates, in one-dimensional case, some realizations from this Gaussian process for different values of W. Smaller the W is, the more rapidly process can vary as a function of location x in the horizontal axis.

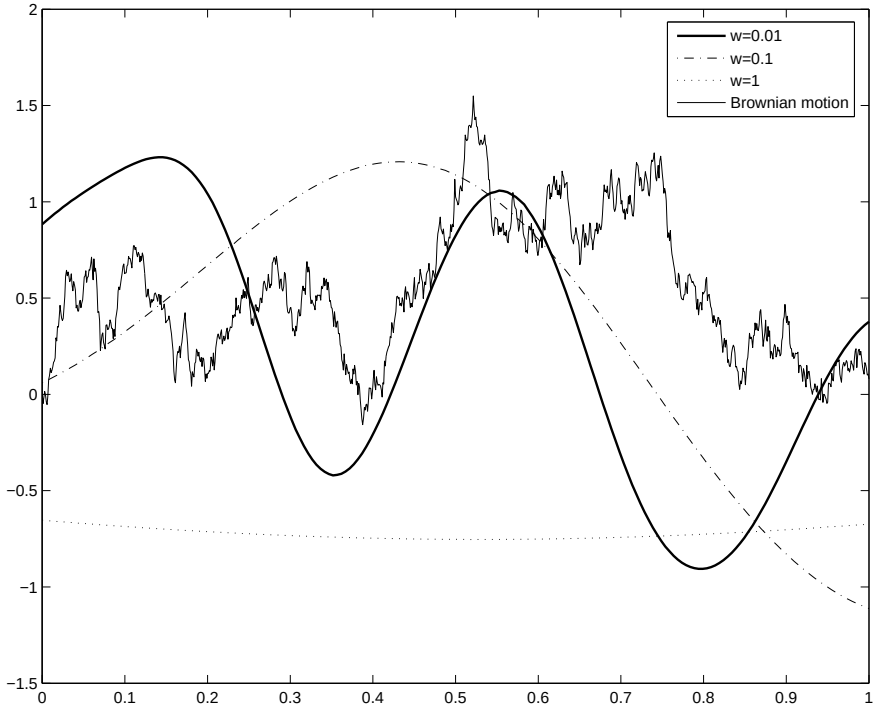

Figure 3.1. Examples of realizations (paths/fields) of Gaussian processes: w refers to Squared exponential covariance function with parameters $W = w$ and $v = 1$, and the latter one is realization of Brownian motion

As we are going to make use of the Squared exponential covariance, it is proper to show that it is a true covariance function, that is a non-negative definite function. This can be done using the *spectral representation* of C. It is immediately evident that the Squared exponential can be written as function of single argument $x_i - x_j$. Gaussian processes with such covariance function (and constant mean function) are *stationary*, that is, simultaneous translation of indices does not alter distribution. For such covariances we can denote $C(x_i, x_j) =: C(x_i - x_j)$, and following theorem holds.

Theorem 3.4 (Spectral representation of covariance function)**.** *A continuous*

complex-valued function $C : \mathbb{R}^d \to \mathbb{C}$ is non-negative definite (i.e. covariance function) if and only if there exists a finite measure ν ($\nu(\mathbb{R}^d) < \infty$) on the Borel-algebra $\mathcal{B}^d$ such that

$$C(x) = \int_{\mathbb{R}^d} e^{i\lambda^T x} \nu(d\lambda), \qquad x \in \mathbb{R}^d, i = \sqrt{-1}. \tag{3.5}$$

Proof. Statement of the theorem is adopted from Adler & Taylor, (2007), but very general, english-language, proof should be found e.g. from Hewitt & Ross, (1997), p. 293, theorem 33.3.

□

If we now set ν to be a centered Gaussian measure (multiplied with constant v) with covariance W^{-1}, application of Gaussian characteristic function equation 2.2 shows that C in above equation 3.5 corresponds to the Squared exponential covariance function. Hence, it follows from the Spectral representation theorem that "Squared exponential covariance function" is non-negative definite, that is, a proper covariance function (as anticipated from the naming).

This Gaussian process model does not yet incorporate any information from the actual observed data. It can be thought of as a prior model. To describe how Y changes as a function of any value x, we will use theorem 3.3. First, let

$$\Sigma = \begin{pmatrix} C(x_1, x_1) & C(x_1, x_2) & \dots & C(x_1, x_n) \\ C(x_2, x_1) & C(x_2, x_2) & \dots & C(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ C(x_n, x_1) & C(x_n, x_2) & \dots & C(x_n, x_n) \end{pmatrix}. \tag{3.6}$$

be the covariance of the outcome observations $Y_{data} = (y_1, \dots, y_n)^T$ given the corresponding input observations $X_{data} = (x_1, \dots, x_n)$. Then, assuming the gaussian process model, the joint distribution of observations and a *new* observation (Y, x) will be of the form

$$\begin{pmatrix} Y_{data} \\ Y \end{pmatrix} \sim N_{n+1}\left(0, \begin{pmatrix} \Sigma & C(x_1 : x_n, x) \\ C(x, x_1 : x_n) & C(x, x) \end{pmatrix}\right), \tag{3.7}$$

where $C(x_1 : x_n, x) = (C(x_1, x), \dots, C(x_n, x))^T$ and $C(x, x_1 : x_n) = C(x_1 : x_n, x)^T$. Thus, we can apply Theorem 3.3 to get the conditional distribution of new observation at x given the old ones. Furthermore, we can *choose* the coordinate-values (index/"position") x for the new value. In essence, we get the distribution of Y in any point x, conditioned for the observed data. Due to Gaussianity, this is equivalent to solving the conditional expectation and covariance. Hence, we have a map $x \mapsto \varphi_{Y_x|(Y_{data}, X_{data})}$, that is, a map from predictor space to conditional distributions of outcomes.

One does not usually observe any direct function f of X, but a noisy version of it. This is why regression problems are typically formulated so that outcome is a function of input with added (Gaussian) noise realization ξ, that is $Y =$

$f(X) + \Xi$. However, since Ξ is independent of $f(X)$ its contribution is easily accounted by adding a multiplicative of indicator function, $(x, x') \mapsto \sigma_\xi 1_0(x - x')$, to covariance function of observations (this equals σ_ξ if $x = x'$, and 0 otherwise). In essence, instead of Σ, we consider $K := (\Sigma + \sigma_\xi I_n)$, where scalar σ_ξ estimates the noise variance. Without further mentioning, we will always make this modification to the covariance function. Then, the joint distribution of data and the new value is.

$$\begin{pmatrix} Y_{data} \\ f(x) \end{pmatrix} \sim N_{n+1}\left(0, \begin{pmatrix} K & C(x_1 : x_n, x) \\ C(x, x_1 : x_n) & C(x, x) \end{pmatrix}\right). \tag{3.8}$$

By using theorem 3.3, we arrive to predicting distribution for $f(x)$, which is Gaussian, and thus defined by mean and variance

$$\hat{f}(x) := E[f(x)|Y_{data}, X_{data}] = C(x_1 : x_n, x)K^{-1}Y_{data}, \tag{3.9}$$

$$\widehat{Var}[f(x)] := C(x, x) - C(x, x_1 : x_n)K^{-1}C(x_1 : x_n, x). \tag{3.10}$$

Notice that $\widehat{Var}[f(x)]$ refers to conditional uncertainty regarding to which event $\omega \in \Omega$ applies, giving particular $f(x) = f(x, \omega)$. This uncertainty stems from our lack of knowledge about the true latent function, and is a decreasing function of n (provided that observations sample near the relevant area of index-space). If we are interested about the uncertainty in possible future *observation*, $Y_x(\omega) = f(x, \omega) + \Xi(\omega)$, then $\sigma_\xi = \widehat{Var}[\Xi]$ must be added to $\widehat{Var}[f(x)]$ to take into account the independent noise in each observed Y_x. This uncertainty does not decrease with increasing amount of observations. Had we assumed a priori that $E[Y_x] = g(x)$ for some function g, instead of $E[Y_x] \equiv 0$, we would see that equation 3.9 is replaced with

$$\hat{f}(x) - g(x) = C(x_1 : x_n, x)K^{-1}\left(Y_{data} - (g(x_1, \ldots, g(x_n))^T\right),$$

meaning that we are doing linear inference based on the residuals of prior assumption. Because this scaling is independent of Y_{data} it is most convenient to set $g \equiv 0$.

This far, we have assumed that we know the correct parameters for covariance matrix $K =: K(v, W, \sigma_\xi) =: K(\theta)$, collectively referred as $\theta := (v, W, \sigma_\xi)$. This rarely is the case. Instead, standard Maximum likelihood estimation can be applied to learn them from the data. The observed outcome data is distributed as $N(0, K(\theta))$, with the density $\varphi(Y_{data}|X_{data}; \theta)$. The value of the probability density function (pdf) of observed data, $\varphi(y_{1:n}|x_{1:n}; \theta)$, can be viewed as a function of θ. This is called the *likelihood function*, and often denoted as $L(\theta)$. It, or usually it's logarithm, can be maximized to find the "most likely" value of θ. Conjugate gradient-based optimization routine works well, as most computational cost per iteration comes from inverting K (Rasmussen & Williams, 2006, p. 114-115). Since logarithm is a monotonic function

$$\text{argmax}_\theta L(\theta) = \text{argmax}_\theta \log L(\theta),$$

and one typically maximizes a simpler function $\ell(\theta) := \log L(\theta)$, instead of $L(\theta)$. In this case,

$$\begin{aligned}\ell(\theta) &= \log \varphi(Y_{data}|X_{data};\theta) \\ &= \log\left(\frac{1}{(2\pi)^{n/2}|K(\theta)|^{1/2}} e^{-\frac{1}{2}Y_{data}^T K(\theta)^{-1} Y_{data}}\right) \\ &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|K(\theta)| - \frac{1}{2}Y_{data}^T K(\theta)^{-1} Y_{data}. \qquad (3.11)\end{aligned}$$

We skip these statistical calculations for now, as we have not yet arrived to our desired model. Let us just state that optimization task is feasible. This maximization procedure is integral part of the *learning* of latent function f from the observations, as it also (in addition to Gaussian conditioning), determines the probability measure of the function space.

Figure 3.2 shows one-dimensional demonstration about the predictive distribution of $f(x) = f(x,\omega)$. We first took 30 data points (open circles) from the realization of the figure 3.1 ($w = 0.01$), and added Gaussian noise with standard deviation 1/2 (stars). After this, we acted as if we only knew the data points denoted by the stars in figure. Then, the likelihood $\varphi(Y_{data}|X_{data},\theta)$ was maximized with a gradient-based optimization routine, and conditional distribution given these 30 data points (stars) was calculated for points 'x' in the horizontal axis of the figure. Solid line in the figure shows estimated $\hat{f}(x)$, and grey area corresponds to points in the confidence interval $(\hat{f}(x) - 2\sqrt{\widehat{Var}[f(x)]}, \hat{f}(x) + 2\sqrt{\widehat{Var}[f(x)]})$, that is to points within two standard deviations from the mean. From this figure we can already anticipate that, in spite of its simplicity, GPR may be a powerful data-driven approach to regression also in larger dimensions. It can automatically find, or "learn", a function from very large class of functions, given some noisy observations from this function. Indeed, it has been shown to perform better, or as well as, many known methods (Rasmussen, 1996). We can also give a direct visual proof of the capacity of GPR in finding hidden functions amidst the noise, as follows.

Consider two functions on a real-plane, $f(x,y) := (x+y)^2 + 2x$ and $g(x,y) := \sin(2x) + y^2$, observed for evenly distributed grid of 15-by-15 observation within $[-2,2]^{\times 2} \subset \mathbb{R}^d$, as in the left column of figure 3. Let us then consider a more realistic situation where we observe only $Y_{(x,z)} := f(x,z) + 5\xi$ and $U_{(x,y)} := g(x,y) + 2\xi$, where $\xi \sim N_1(0,1)$ is a Gaussian perturbation/error (middle column). Furthermore, consider that we have no knowledge about the form of f or g, nor from the multiplicative constant of the random ξ. Again, we assumed the above discussed GPR model, maximized the likelihood, and made a prediction $\hat{f}(x,z) = E[f(x,z)|Y_{data},(x_{data},z_{data})]$, and similarly for $\hat{g}(x,y)$ (right column). While functions are not reproduced perfectly, it is obvious that this is a good result. Visual inspection of the middle columns of figure 3 could not hint this result, and human visual system is fairly powerful pattern detector in sufficiently low dimension. While we neglected the actual likelihood maximization

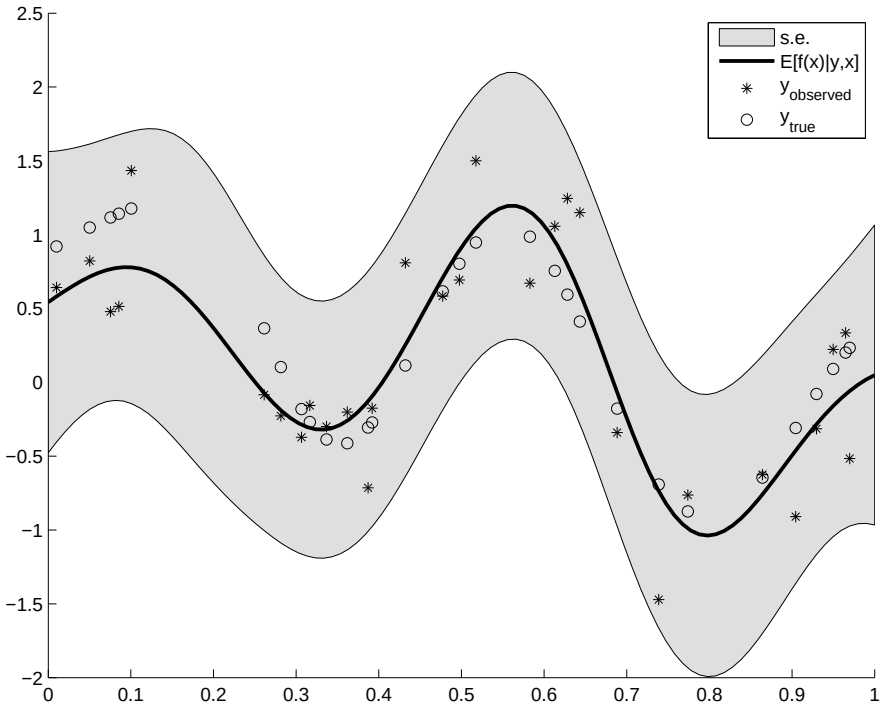

Figure 3.2. Mean and two standard deviations of Conditional Gaussian process given the observations (stars)

for now, we will return to it later on. Also, next section will assess theoretical arguments for the surprising power of this fairly simple smoothing device.

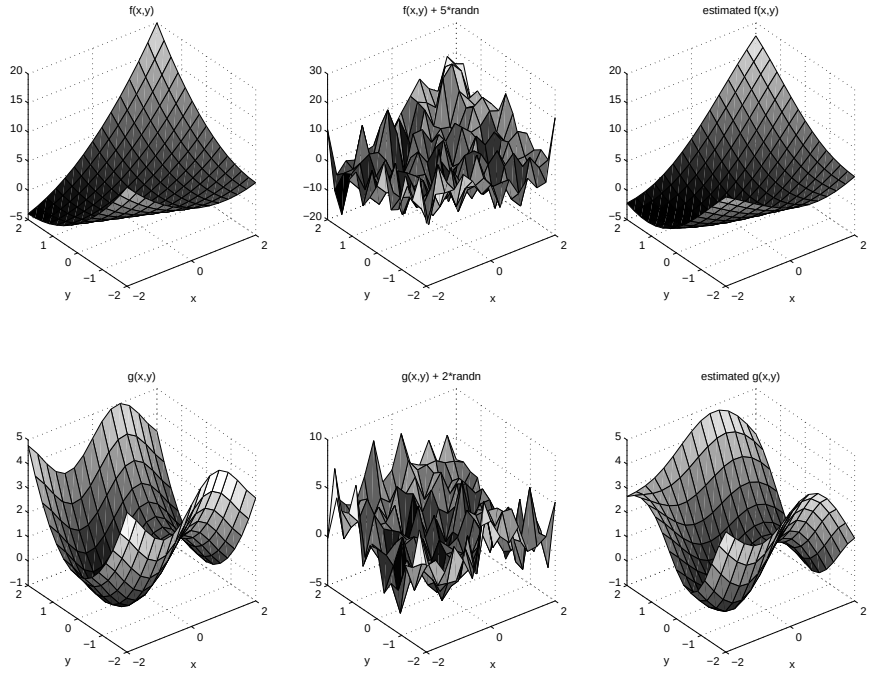

Figure 3.3. Denoising functions on plane using GPR

3.3 Theoretical perspective

In this chapter we attempt to build some intuition for Gaussian processes in regression context. Origins of GPR lie in the Geostatistics, where it is known as Kriging. Gaussian random field can be used, for example, to estimate concentration of some mineral beneath the ground, given some amount of measured sample sites. Given above examples, it can be viewed as more general regression tool as well. But why do things work out so neatly for GPR? Is there a way to better understand why maximization of the likelihood function φ so readily leads to right covariance function, allowing us to find the desired function value with a simple linear smoothing (matrix product) on the observed outcome values of Gaussian process? Not so surprisingly, the answer is yes. But, we need some tools to provide this reasoning. Actually, there are several ways to gain intuition, but let us begin with a combination of constructs known as *Reproducing kernel Hilbert space* (RKHS) and *Bayesian statistics*. After this we will shortly present a connection between Markov processes and a subset of Gaussian processes. In general, this section provides broader perspective and is not absolutely necessary for understanding the rest of current work.

Some notions from Functional analysis are unavoidable. Recall that *Hillbert space*, is a *Banach space* that allows an inner product. A Banach space is a *complete* vector space with a norm $\| \cdot \|$. A metric space is complete when every *Cauchy sequence* converges to a point *in* that space. If a sequence $(x_i)_{i=1}^{\infty}$ satisfies that, for every $\varepsilon > 0$, there is an integer n such that $\| x_i - x_j \| < \varepsilon$ for all $i, j \geq n$, it is called a Cauchy sequence. If V is a vector space, inner product is a map $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ which is linear, symmetric in its arguments, and positive definite, that is, $\langle x, x \rangle > 0$ for all $x \neq 0$. The inner product defines a norm with $\| x \|_2 := \sqrt{\langle x, x \rangle}$.

RKHS is constructed from the covariance function $C(\cdot, \cdot)$ of a Gaussian process. Let T be some index set, for example $T \subset \mathbb{R}^d$. We follow the exposition in Adler & Taylor, (2007), and start with a following function space:

$$S = \{u : T \to \mathbb{R} : u(\cdot) = \sum_{i=1}^{n} a_i C(x_i, \cdot), a_i \in \mathbb{R}, x_i \in T\}.$$

This space allows an inner product defined by

$$(u, v)_H = \left(\sum_{i=1}^{n} a_i C(x_i, \cdot), \sum_{j=1}^{m} b_i C(x_j, \cdot) \right)_H := \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j C(x_i, x_j),$$

if C is positive definite covariance function. Furthermore, for fixed $x \in T$, if we consider $C(x, \cdot)$ as a function of its second argument, $(\cdot, \cdot)_H$ satisfies an unusual *reproducing kernel* property:

$$(u, C(x, \cdot))_H = \left(\sum_{i=1}^{n} a_i C(x_i, \cdot), C(x, \cdot) \right)_H = \sum_{i=1}^{n} a_i C(x_i, x) = u(x).$$

Inner product $(\cdot, \cdot)_H$ defines a norm by $\| u \|_H := (u, u)_H^{1/2}$. Closure of S under this norm is a Hilbert space of real-valued functions, denoted as $H(C)$, and defined by the covariance function C. This is the RKHS. A Hilbert space is said to have an *orthonormal basis*, $(e_i)_{i=1}^{\infty}, e_i \in V$ if all its elements, u, can be expressed as

$$u = \sum_{i=1}^{\infty} \langle u, e_i \rangle e_i,$$

where $\| e_i \|_2 = \langle e_i, e_i \rangle^{1/2} = 1$ for all i, and $\langle e_i, e_j \rangle = 0$ for all $i \neq j$. While detailed construction is out of the scope here, *Mercer's theorem* guarantees that $H(C)$ has an orthonormal basis (Adler & Taylor, 2007; Rasmussen & Williams, 2006).

Consider then a space defined as $\mathcal{H} := \text{span}\{Y_x, x \in T\}$, that is, countable linear combinations of the values of centered Gaussian process in some indices. This space inherits the usual L^2-inner product, $\langle X, Y \rangle := E[XY]$, discussed in the Introduction. Now, let us define a linear map $\Gamma : S \to \mathcal{H}$,

$$\Gamma(u) = \Gamma \left(\sum_{i=1}^{n} a_i C(x_i, \cdot) \right) := \sum_{i=1}^{n} a_i Y_{x_i}.$$

Clearly, as a sum of Gaussian variables, $\Gamma(u)$ is Gaussian. It also defines a (norm-preserving) linear isomorphism:

$$\begin{aligned} \| \Gamma \left(\sum_{i=1}^{n} a_i C(x_i, \cdot) \right) \|_2^2 &= | \sum_{i=1}^{n} \sum_{j=1}^{n} a_i E[Y_{x_i} Y_{x_j}] a_j | \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i C(x_i, x_j) a_j \\ &= \| u \|_H^2 . \end{aligned}$$

Consequently, Γ extends to all of $H(C)$, as is known from functional analysis. Furthermore, all limits are Gaussian, as shown in the proof of the Theorem 2.2 regarding existence of Gaussian process.

If now $(e_i)_{i=1}^{\infty}$ is an orthonormal basis of $H(C)$, setting $\xi_i = \Gamma(e_i)$ for all i yields an orthonormal basis for $\mathcal{H}$. We must have that all ξ_i are Gaussian. Since $\| \xi_i \|_2 = 1$, $\langle \xi_i, \xi_j \rangle = E[\xi_i \xi_j] = 0$, and ξ_i and ξ_j are Gaussian, it follows that $E[\xi_i \xi_j] = E[\xi_i] E[\xi_j] = 0$ for all $i \neq j$, and further that all ξ_i have distribution $N_1(0, 1)$. Also,

$$Y_x = \sum_{i=1}^{\infty} \xi_i E[Y_x \xi_i],$$

where the sum converges in L^2. Because Γ was isometry with $\Gamma^{-1}(Y_x) = C(x, \cdot)$, and $(\cdot, \cdot)_H$ has a reproducing kernel property,

$$E[Y_x \xi_i] = (C(x, \cdot), e_i)_H = e_i(x).$$

Collecting above results establishes following theorem, which was the purpose of current functional analytic discussion:

Theorem 3.5 (Orthogonal expansion of Gaussian process)**.** *If the sequence of real-valued functions, $(e_i)_{i=1}^{\infty}$, is an orthonormal basis for $H(C)$, then zero mean Gaussian process $\{Y_x\}$ with covariance C has the L^2-representation*

$$Y_x = \sum_{i=1}^{\infty} \xi_i e_i(x),$$

where $(\xi_i)_{i=1}^{\infty}$ is orthonormal sequence of centered Gaussian variables given by $\Gamma(e_i)$.

This theorem also allows one to show, via tail-event argument, that

Theorem 3.6. *$x \mapsto Y_x$ is almost surely continuous if, and only if, the sum in theorem 3.5 converges uniformly in T with probability* 1*.*

Proof. See e.g. chapter 3 in Adler & Taylor, (2007).

□

Hence, Gaussian processes almost surely are either continuous or discontinuous, without middle ground. Also, the roughness of the process is controlled by basis functions of $H(C)$, thus, ultimately by the covariance function C.

Theorem 3.5 opens up an interpretation, previously presented (at least) in Rasmussen & Williams, (2006). Recall that linear regression model with one variable can be defined as $Y_x = \beta_1 x + \xi$. Familiar extension of this standard model to cover nonlinearities with respect to x is to seek estimate for $Y_x \approx \sum_{i=1}^{n} \beta_i e_i(x)$, where $(e_i)_{i=1}^{n}$ is some set of nonlinear *basis functions*. If we now replace fixed set of coefficients $(\beta_i)_{i=1}^{n}$ with a random vector $(\xi_i)_{i=1}^{n} =: \beta$, we have given a Bayesian *prior distribution* to weight of each basis function in Y. Let (Y, x) denote some observed data $((Y_{x_1}, x_1), \ldots, (Y_{x_m}, x_m))$. According to a well-known Bayes theorem, in the case that we have densities $p_x(\cdot)$ for all relevant distributions,

$$p_x(\beta|Y) = \frac{p_x(Y|\beta)p_x(\beta)}{\int p_x(Y|\beta)p_x(\beta)d\beta} = \frac{p_x(Y|\beta)p_x(\beta)}{p_x(Y)}, \tag{3.12}$$

is the posterior probability of regression coefficients (equivalently basis function weights) given the observed data. We wrote x in the subindices of densities, p_x, to remind that they depend on constant x (or alternatively, on fully independent "marginal model" $p_X(x)$, as often interpreted in statistics).

Quantity $p_x(Y)$ in 3.12 is known as the *evidence* (or marginal distribution) of the data, and it plays an important role in Bayesian model selection. Notice that it involves an integral of data likelihood with respect to prior measure $p_x(\beta)d\beta$. This automatically implements what is know as the *Occam's razor* in philosophy of science (MacKay, 2003, chap. 28). It means that one should favor a simple hypothesis over a complex one, when it suffices to explain the observed data. If one defines a prior that sets probability mass to a wide volume of parameter space (diffuse/uninformative prior) then constraints $\int p_x(\beta)d\beta = 1$ and $p_x(\beta) \geq 0$ imply that $p_x(\beta)$, and hence also $p_x(Y)$, must have mostly small values. If, on the other hand, $p_x(\beta)d\beta$ puts mass only to very small set, and it does not happen to be near the maximum of likelihood $p_x(Y|\beta)$ for the current observations, also this leads to small evidence. Optimal situation occurs with a "small" model (simple hypothesis) setting probability mass to correct region of parameter space (and to correct *dimensional* parameter space). Notice further, that $p_x(Y)$, being function of data only, involves no more parameters β. In fact, it is a function of data *and* chosen model, represented by likelihood and prior structures.

Drawing a direct analogy from above discussion to GPR, let the sum $\sum_{i=1}^{n} \xi_i e_i(x)$ tend to infinity as in theorem 3.5. The space of possible model structures is now indexed by the parameters θ defining the precise form of covariance function C. Due to nice analytic properties of Gaussian families, we can make all the relevant predictions without never explicitly writing out ξ_i or $e_i(x)$. When we maximize the likelihood $\varphi(Y_{data}|X_{data}, K(\theta))$ with respect to theta, we are actually maximizing the evidence of model, not just likelihood. "Parameters" of the orthogonal expansion are just marginalized away from the visible equations.

Hence, in spite of the fact that Gaussian processes can model very many functions, we are not overfitting to data, as would otherwise happen with infinite predicting functions. That is, the Occam's razor is built-in to GPR. Of course, this goes only so far as we may regard that we have chosen a right covariance function form, that is, right mapping $\theta \mapsto C_\theta(\cdot,\cdot)$. In any case, this should offer a strong intuition for why GPR works as well as it does. In chapter six we will see that, in the case of the Squared exponential covariance function, basis functions of $H(C)$ *and* L^2-expansion (given by Mercer's theorem) are Gaussian kernels centered densely[1] in T.

When K is the covariance function for n observations as in previous section, log-likelihood in GPR-model decomposes as

$$\ell(\theta) = -\frac{1}{2}Y_{data}^T K(\theta)^{-1} Y_{data} - \frac{1}{2}\log|K(\theta)| - \frac{n}{2}\log(2\pi).$$

In the light of above discussion, these parts of a sum have clear interpretation. The first, and only, term depending on outcome observations is $-\frac{1}{2}Y_{data}^T K(\theta)^{-1} Y_{data}$. This is the measure for data-fit, that is, for how well the model capture properties of current data. Second term, $-\frac{1}{2}\log|K(\theta)|$, is a *complexity penalty*, which depends only on the covariance function and indexing observations. Recall, that absolute value of determinant gives the *volume* of (hyper) parellelepiped formed by row vectors of the matrix, and that covariance matrix is a linear description for how random vector varies in the multidimensional space. Hence, the more degrees of freedom we leave for the data ("larger the volume of covariance matrix"), larger is the complexity penalty. Hence, in optimization of the likelihood, there is a trade-off between the data-fit and allowed complexity. Last term, $-\frac{n}{2}\log(2\pi)$, is a normalization constant for probability density.

While orthogonal expansions and Bayesian regression offer very general framework for understanding GPR, in somewhat more restricted context, useful insight may stem from a research line known as the Dynkins program (Ylvisaker, 1987). This approach connects Gaussian processes with much investigated Markov processes, leading to *Gaussian Markov-associated processes* (G-MAP). We say that

Definition 3.3. *Gaussian process $\{Y_x\}$ is a G-MAP on T if for any $x \in T$ and finite subset U of T,*

$$E[Y_x | Y_{x_i}, x_i \in U] =: E^U[Y_x] = \sum_i p_x^U(x_i) Y_i, \tag{3.13}$$

where $p_x^U(x_i) \geq 0$ form a sub-distribution on U, that is, $\sum_i p_x^U(x_i) \leq 1$.

Note that in predictive equation 3.9 we already have that expected value of the process is a linear combination of observations, as in above, although in above, observations are considered as random (not yet observed), and $E^U[Y_x]$

[1] A set $U \subset T$ is dense in T if its closure equals T.

is thus a random variable. Equation 3.9 does not necessarily define a sub-distribution on observations. In fact, G-MAPs are non-smooth processes (Ylvisaker, 1987), excluding covariances such as the Squared exponential. Nonetheless, they have a property which make them interesting for questions of designing experiments, which is why we very briefly mention them here, as a potential future research direction. Current work does not go farther to this direction as we wish to retain connection with the squared exponential covariance.

In some cases it may happen that collecting indexing observations (behavioral data, questionnaires, etc.) is "cheap" in comparison to collecting outcome observations. For example, it is often of interest in medicine to associate some behavior to neural transmitter concentrations in human brain. Drugs that attempt to change behavior and mood often target these transmitters, but there are lot of open questions for research. Measurement of transmitter concentration may be invasive, involving procedures such as drawing bone marrow or injecting radioactive tracer to blood stream to aid brain imaging. In such research context, one wishes to expose as few people as possible to procedure, while still collecting representative data that allows statistical inference. An optimal research *design* might then mean the best sampling of indexing space T. In the context of Gaussian processes, we might wish to minimize our uncertainty regarding to true (expected) function in some interesting subset of indices T. That is, we wish to find out which indexing observations minimize predictive variance of equation 3.10, *before* we have actually collected any observations. It is then possible to collect large set of indexing observations and choose only those individuals/indices to further outcome measurement who offer reliable estimate with small amount of observations. If, in addition, we wish to use a covariance that leads to G-MAP, some answers may be found via the Dynkin's finding that for each G-MAP, there exists an associated Markov-process.

For simplicity, we consider a countable index-space T. By definition, one can then arrange T to some increasing order with bijection to natural numbers $\mathbb{N}$ (note that high-dimensional grid of arbitrary precision is countable). Let us now denote indices simply with natural numbers. A stochastic process $\{Z\}$ is said to be Markov-chain if distribution of next observation depends only on the one before it, that is

$$P\left(Z_{n+1} \in A | (Z_1, Z_2, \ldots, Z_n)\right) = P(Z_{n+1} \in A | Z_n)$$

for all sets $A \in \mathcal{F}$. Due to this simplifying property, Markov chains are easy to analyze and simulate. If $\{Z_n\}_{n\in\mathbb{N}}$ is a Markov process one may define a *transition probability* $p(x, y) = P(Z_1 = y | Z_0 = x)$ which is a probability that chain moves to y in one step, starting from x. Then, due to defining Markov-property, probability of moving to y in *n steps* is $p^n(x, y) = \sum_{z\in T} p^{n-1}(x, z)p(z, y)$. Let then $N(y) := \sum_{n=0}^{\infty} 1_{\{Z_n=y\}}$ be the total number of visits of Z to y. We may now define the *Green function* of Z as

$$G(x, y) = E_x[N(y)] = \sum_{n=0}^{\infty} p^n(x, y), \tag{3.14}$$

where x in the sub-index of expectation operator refers to starting point $Z_0 = x$. If $G(x, y)$ is finite, Z is expected to visit y only finite number of times (process is *transient*). If one wishes to examine non-transient (recurrent) process there is a technical trick that one may force an additional imaginary observation that is defined as a *killing* state. That is, after entering this state, process is no longer in T. It turns out that $G : T \times T \to \mathbb{R}$ is non-negative definite, and hence there is a zero mean Gaussian process $\{Y_x\}_{x \in T}$ (Ylvisaker, 1987).

Theorem 3.7. *The Gaussian process $\{Y_x\}_{x \in T}$ with covariance function G given by equation 3.14 is a G-MAP on T, and $\{p_x^U(x_i)\}_i$ is the first-hit distribution, $\{P_x(\inf_{Z_n \in U, n>0} Z_n = \{x_i\})\}_i$, of associated Markov-process Z on U, starting from index-state x.*

□

If transient case is excluded, sub-distribution in the definition of G-MAP will be a proper probability distribution (Ylvisaker, 1987). If we now define the *error process*

$$(Y - E^U[Y])_x := Y_x - E^U[Y_x],$$

via definition of G-MAP, we get error process covariance

$$\begin{aligned}
&E[(Y - E^U[Y])_x)(Y - E^U[Y])_x] \\
&\quad = E[Y_x Y_x] - 2E[Y_x E^U[Y_x]] + E[E^U[Y_x] E^U[Y_x]] \\
&\quad = G(x,x) - 2\sum_i p_x^U(x_i) E[Y_x Y_{x_i}] + \sum_i \sum_j p_x^U(x_i) p_x^U(x_j) E[Y_{x_i} Y_{x_j}] \\
&\quad = G(x,x) - 2\sum_i p_x^U(x_i) G(x, x_i) + \sum_i \sum_j p_x^U(x_i) p_x^U(x_j) G(x_i, x_j) \\
&\quad = G(x,x) - \sum_i p_x^U(x_i) G(x_i, x) \\
&\quad = E[\text{number of visits of Z to x starting from x}] \\
&\qquad\quad - E[\text{number of visits of Z to x after hitting U}].
\end{aligned}$$

Error process covariance would equal predictive covariance (eq. 3.10), if $Y_x = f(x)$, that is, f be observed without noise. Thus, we get the uncertainty regarding latent function value at x given observations in the indices $U \subset T$, in terms of expected returns of Markov chain to x. In addition to illuminating interpretation, this opens up a possibility of inferring design questions via Markov chain simulation. Note, however, that these interpretations apply only to a restricted set of covariance functions. We conclude our theoretical section with these prospects for future investigation. For example, Marcus & Rosen, (2006) offer a recent reference on the theoretical connections between Markov and Gaussian processes.

3.4 Problem of measurement error

Above presented regression approach works independently of whether X_{data} is drawn with randomness or by design. Of course, we get information from

the behavior of f mainly in the neighborhood of observations, but sampling distribution need not be explicitly incorporated to practical modeling considerations. However, what happens if we observe X_{data} inaccurately due to some noise source in the measurement procedure of X itself? Consider the classical measurement error model, where $X_{observed} = X_{true} + \xi_x$ and $\xi_x \sim N(0, \sigma_x)$ (Carroll, Ruppert, Stefanski & Crainiceanu, 2006, chap. 1). Depending on the "true" latent function f, $f(x_{observed})$ may be very different from $f(x_{true})$. If we go on forcing the same model, observe that

$$\begin{aligned} Y &= f(X_{observed}) + \xi \\ &= f(X_{true}) - f(X_{true}) + f(X_{observed}) + \xi \\ &=: f(X_{true}) + \xi', \end{aligned}$$

where $\xi' = f(X_{true} + \xi_x) - f(X_{true}) + \xi$ is no longer Normally distributed independent error, but depends from f, X_{true}, ξ_x and ξ. As ξ is independent of other factors in the new error term, $Var[\xi'] \geq Var[\xi]$. Depending on the situation, $Var[\xi']$ may be a lot larger than $Var[\xi]$. Hence, likely consequence of using standard GPR-model with erranously observed input is an inflation of i.i.d. noise variance parameter, σ_ξ. If estimate of i.i.d. noise component is inflated, this means that only small part of observed variance in the data is interpreted as coming from the latent function f, that is, estimate of $\hat{f}$ given by equation 3.9 tends toward constant function. Computer simulations verify that this indeed happens (Dallaire, Besse & Chaib-draa, 2009). Uncertainty estimate $\hat{Var}(f)$ of equation 3.10 may also be overly optimistic for the constant function interpretation. Thus, disregarding noise in indexing observations may lead to high confidence for wrong result.

From above considerations, it follows that we need to somehow take into account ξ_x also. To do this, we need information about it. To gain information, we need to create a model for the measurement procedure and associated error. In this situation, a modeler needs to be familiar with the properties of the actual data that is modeled. For different situations, there are different ways to model the error in predictor variable (Carroll, Ruppert, Stefanski & Crainiceanu, 2006). In the next chapter, we introduce a model for our purpose. Only after this, we return to problem of GPR with uncertain input.

4 Measurement model

4.1 Behavioral scales

Behavioral scale is defined to be a function of a set of items. Typically, these items are answers to questions of a questionnaire, given as numerical values. Let us denote these questions as $q = (q^{(1)}, q^{(2)}, \ldots, q^{(m)})^T$. For example, $q_i^{(j)}$ could stand for the integer value from 1 to 5, depending on whether the individual i thinks his/her behavior agrees well with the description of the question j or not. In typical Likert-scale (coding standard), an individual sets value 1 for $q_i^{(j)}$ if the description does not fit at all to him/her, and value 5 if it fits very well. Sometimes q already stands for the sum of values for several questions. In any case, it is commonly agreed in behavioral sciences that answers to single question, or few of them, are not reliable indicators of a long-term behavior. For this reason, one often wishes to collect answers to very many question (typically c. 50-300), and then to reduce this information to few particularly informative variables, or dimensions. First of all, one wishes to describe a general behavioral tendency, and people are not very able to objectively describe themselves in relation to other population. However, it is much easier to extract objective responses to questions about how one behaves in situations occurring on daily basis (or often enough to be well recollected). One can then derive more general constructions from this information.

Another reason to favor dimension reduction methods, is that they reduce error variation from the measure of interest, given that one has several indicators for it. Essentially, this is a consequence of the Central limit theorem of probability theory. If one has several independent noisy measurements $(q^{(1)}, q^{(2)}, \ldots, q^{(m)})$ from the same common underlying (latent) variable, then their average tends toward the value of this latent variable. Larger the number of measurements/items, m, the less will this average deviate from the latent variable of interest. In fact, the constructed scale often is just a simple average of several items.

In the next section, we will describe a statistical formalism that is a linear description of behavioral scales. This suffices for the purposes of present study, as non-linear functions from items to behavioral scales are virtually non-existent in the literature. It would also be very difficult to come by with this kind of a scale. Usually there is not much more theoretical information, than a knowledge that given item associates with certain underlying construct/latent variable. Data-driven estimation of nonlinearities would be a very challenging

task, given the amount of questions and lack of their precision.

4.2 Standard statistical model - Factor analysis

Standard statistical approach for the modeling of behavioral scales is based on the Factor analysis method (Lawley & Maxwell, 1971; Cudeck & MacCallum, 2007; Tarkkonen & Vehkalahti, 2005). Assume one has m items, $q = (q^{(1)}, q^{(2)}, \ldots, q^{(m)})^T$, which supposedly reflect variation from d latent variables that are elements in a vector $u \in \mathbb{R}^d$. Then, according to Factor analysis model, item values for each individual i are independently generated as

$$q_i = \Lambda u_i + \xi_i, \quad \xi_i \sim N_m(0, \Psi), \quad u_i \sim N_d(0, \Phi). \tag{4.1}$$

Here, $\Lambda \in \mathbb{R}^{m \times d}$ is a fixed matrix of coefficients, or loadings, of u to items q. It tells how much of the true variation of elements of u each item $q^{(j)}$ preserves. Ψ is a diagonal error covariance matrix, and Φ is the covariance of latent variables. The same model is assumed to hold for all individuals. Expectation of u could be non-zero, but this does not yield much more generality for the model, as one can always subtract the mean from observations of q. Since behavioral measures can mainly be interpreted only relative to other individuals, mean values by themselves offer little information. Thus, we prefer to keep notation transparent, and set $E[U] = 0$. From the properties of expectation and linear algebra,

$$Cov[Q] = E[QQ^T] = \Lambda\Phi\Lambda^T + \Psi. \tag{4.2}$$

Estimation of the matrices Λ, Φ and Ψ from the data is not a trivial task, but there are standard routines implemented to nearly every statistical program. Typically, estimation is divided to two parts, because parameters that maximize the likelihood are not unique. Thus, one *extracts* some parameters by maximizing the likelihood, and *rotates* within the set of maximizing parameters according to some criterion (Lawley & Maxwell, 1971; Jennrich, 2007). To see where this terminology comes from, consider the case of *orthogonal*, or uncorrelated, latent variables, i.e. $\Phi = I_d$ (due to Gaussian assumption, this also implies independence of latent variables). In this case, if Λ_0 and Ψ are maximum likelihood estimates, and M is an arbitrary d-by-d rotation matrix (that is, $det(M) = 1$ and $M^{-1} = M^T$), then $\Lambda_1 := \Lambda_0 M$ and Ψ are also maximum likelihood estimates. This follows from the equality:

$$\Lambda_0\Lambda_0^T + \Psi = \Lambda_0 M M^T \Lambda_0^T + \Psi = \Lambda_1\Lambda_1^T + \Psi.$$

In this case, extraction step refers to finding some Λ_0, and rotation refers to finding matrix M such that Λ_1 becomes, in some sense, convenient, or interpretable. Usually this convenience criterion is some sort of simplicity requirement, setting as many of elements of Λ_1 near zero as possible.

Further information on the topic of rotation can be found from the literature (Lawley & Maxwell, 1971; Jennrich, 2007). In the next chapter, we will show a simple method for the extraction step. In general, more advanced Markov chain Monte Carlo estimation methods, such as Gibbs sampling (Geman & Geman, 1984, , see also estimation methods section in the chapter below), offer distinct benefits compared to more traditional maximum likelihood methods (Lawley & Maxwell, 1971). This is because of so called *Heywood cases* (Martin & McDonald, 1975). Maximum likelihood methods may (fairly often) end up with a solution where some of the diagonal values in Ψ are zeros, or negative. Clearly, it makes no sense for the variance to be negative. It rarely makes sense, that some item $q^{(j)}$ reflects pure latent variation, without any noise, corresponding to case of zero value in Ψ_{jj}. Bayesian estimation avoids these cases by putting a prior distribution for Ψ, which assigns null probability for diagonal values less than, or equal, to zero. For example, Lopes & West, (2004) offer a good implementation for the extraction via Gibbs sampling. This appears to slightly outperform maximum likelihood methods even in a simple ideal data case (personal simulations, not shown). Another good thing that comes with a Gibbs sampler is simulated set of random values from the posterior distribution of latent values $(U_1, \ldots, U_n)$. From this, one gets estimates for the mean and standard error of each individual's true/latent traits. Also, more flexible model assessment becomes possible with a Bayesian approach (Lopes & West, 2004).

Given that it is possible to estimate the Factor analysis model, and to estimate the "true value" u_i for each individual i, one might guess that this is a standard practice. After all, it would, to some extent, solve the problem of measurement error discussed in the end of the preceding chapter. However, Factor analysis model involves plenty of unknown parameters, and also other things (Ellis, 2004), which make it an unstable tool. Construction of a behavioral scale, from initial pondering of appropriate questions to a final tested product, is a huge undertaking. One wishes to use it many times and in many places, in as identical form as possible. Thus, Factor analysis is, in practice, often used mainly to confirm what questions reflect most the given latent variable, and what questions do not. Means to measure u, the scale, is then formed manually, and typically each element of measured u is taken to be just an average of certain subset of items in q. Also, any *estimate* for u_i still contains noise in the form of estimation error, and possibly also via model miss-specification. Basing one complex model estimation (GPR) on estimate from another complicated procedure (Factor analysis) may not yield good results in practice. Hence, we will devote the next chapter for estimation considerations in the case where these models need to be somehow combined. In below, we discuss how to derive error estimates for u_i, further needed in the chapters five and six. We need a basic result, adopted from Cappé et al., (2005).

Proposition 4.1 (Conditioning in the Gaussian Linear Model)**.** *Let U and V be two independent Gaussian random vectors with $E[U] = \mu_U$, $Cov(U) = \Sigma_U$*

and $Cov(V) = \Sigma_V$*. Assume* $E[V] = 0$*, and consider the model*

$$Q = \Lambda U + V,$$

where Λ *is deterministic matrix of appropriate dimensions. Further assume that* $\Lambda\Sigma_U\Lambda^T + \Sigma_V$ *is full-rank matrix. Then*

$$\begin{aligned} E[U|Q] &= E[U] + Cov(U,Q)Cov(Q)^{-1}(Q - E[Q]) \\ &= \mu_U + \Sigma_U\Lambda^T(\Lambda\Sigma_U\Lambda^T + \Sigma_V)^{-1}(Q - \Lambda\mu_U) \end{aligned} \tag{4.3}$$

and

$$\begin{aligned} Cov(U|Q) &= Cov(U - E[U|Q]) = E[(U - E[U|Q])U^T] \\ &= \Sigma_U - \Sigma_U\Lambda^T(\Lambda\Sigma_U\Lambda^T + \Sigma_V)^{-1}\Lambda\Sigma_U. \end{aligned} \tag{4.4}$$

Proof. Denote

$$\hat{U} := E[U] + Cov(U,Q)Cov(Q)^{-1}(Q - E[Q]),$$

which implies

$$Cov(U - \hat{U}, Q) = Cov(U,Q) - Cov(U,Q)Cov(Q)^{-1}Cov(Q) = 0.$$

The random vectors Q and $U - \hat{U}$ are thus jointly Gaussian (as linear transformation of Gaussian random vector, see proof of theorem 3.3) and uncorrelated. Due to Gaussian distribution, uncorrelatedness also implies that they are independent. $\hat{U}$ is $\sigma(Q)$-measurable, as a linear combination of the components of Q. For $\sigma(Q)$-independent variable it holds that $E[U - \hat{U}|Q] = E[U - \hat{U}]$, because for any $A \in \sigma(Q)$, $E[E[U - \hat{U}|Q]1_A] = E[(U - \hat{U})1_A] = E[U - \hat{U}]E[1_A]$, due to independence. Hence,

$$\begin{aligned} E[U|Q] &= E[\hat{U} + (U - \hat{U})|Q] \\ &= \hat{U} + E[(U - \hat{U})] \\ &= \hat{U} + 0 = \hat{U}, \end{aligned}$$

and

$$Cov(U|Q) := E[(U - \hat{U})(U - \hat{U})^T|Q] = E[(U - \hat{U})(U - \hat{U})^T] =: Cov(U - \hat{U}).$$

Finally, we obtain the desired covariance by noting that

$$Cov(U - \hat{U}) = E[(U - \hat{U})(U - \hat{U})^T] = E[(U - \hat{U})U^T],$$

which follows from

$$\begin{aligned} &E[(U - E[U|Q])E[U|Q]^T] \\ &\quad = E[E[UE[U|Q]^T|Q]] - E[E[U|Q]E[U|Q]^T] = 0 \end{aligned}$$

(Tower property). Rest of the statement follows directly from the assumed linear model structure, where e.g.

$$\begin{aligned} Cov(U,Q) &= E[(U - E[U])(Q - E[Q])^T] \\ &= E[(U - E[U])\left(\Lambda(U - E[U]) + V\right)^T] = Cov(U)\Lambda^T, \end{aligned}$$

and so on.

□

If we know values of Λ and have observed those of $Q = q$, we get an estimate for U directly from the above proposition, by taking $E[U|Q = q]$. Similarly, we get estimate for the covariance of estimator $(U|Q = q)$ from this proposition as $Cov(U|Q = q)$.

4.3 Error variance estimate for behavioral scales

Let Υ_i be some estimate for the error of our estimated latent value u_i for individual i. If we have estimated the Factor analysis model via Gibbs sampling, as discussed above, then we get a direct estimate for Υ_i as the posterior covariance of u_i. Or we could use $Cov(U|Q = q)$ as above. This error estimate will come in handy when making predictions with GPR, as will be seen in chapter 6. For now, we discuss more about the case where one wishes to use a standardized scale, derived from items q according to prior research and theory. Here (as in Tarkkonen & Vehkalahti, 2005), we take behavioral scale to be a vector $x := A^T q \in \mathbb{R}^d$, where $q \in \mathbb{R}^m$ are the observed items for an individual, and A^T is the matrix of fixed linear map $A^T : \mathbb{R}^m \to \mathbb{R}^d$. Matrix A describes how the scale for latent $u \in \mathbb{R}^d$ is formed from the observed $q \in \mathbb{R}^m$ $(m > d)$. With appropriate fixed elements of A, most practical behavioral scales in the literature can be described within this generality. For a concrete example, consider that we have 6 items/questions, for which three first ones seem to measure/reflect one latent construct, while the rest measure another. Then, for standard (averaging) scale, A would be

$$A^T = \begin{pmatrix} 1/3 & 1/3 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \end{pmatrix}.$$

While we consider that the scales used in practice are defined by some mapping/matrix A, we still assume that the Factor analysis model itself holds for some parameter values. This is rather reasonable assumption. As discussed, non-linear scales are difficult to come by. As is well-known, Normal/Gaussian distribution is, in many sense, an ideal description of measurement error. Also, complex behavioral traits tend toward Gaussian distribution (Plomin, Haworth & Davis, 2009), as do (according to Central limit theorem) variables that are sums of many separate factors. These assumptions imply also a Gaussian distribution for q, with a covariance given by equation 4.2. As discussed, elements

of q may take only five discrete values, which sometimes is a problem, although rarely too much in practice. Often, elements of q already are sums of many discrete-valued items, and approximate Gaussian variables fairly well. Here, we conform to standard practice and leave this matter unattended. Research is done all the time to overcome this approximative step, but numerical complexity grows rapidly as new unknown parameters must be introduced (Moustaki, 2007).

With these assumptions, observed/measured values of the scale, decompose as $x_i = A^T q_i = A^T \Lambda u_i + A^T \xi_i$, with between-individual covariance

$$Cov[X] = A^T Cov(Q) A = A^T \Lambda \Phi \Lambda^T A + A^T \Psi A. \tag{4.5}$$

In this *covariance decomposition*, $A^T \Psi A$ corresponds to part of the covariance that reflects noise variation, whereas $A^T \Lambda \Phi \Lambda^T A$ is the between-individual latent variance that can be expected to predict something (Tarkkonen & Vehkalahti, 2005). Thus, $A^T \Psi A$ delivers an estimate of the measurement error covariance in X. In addition, we may think that each observed $x_i^{(j)} \in \mathbb{R}$ contains additive noise component with variance $[A^T \Psi A]_{jj}$, for all individuals i. If we do not force any standard scale, proposition 4.1 directly gives us the optimal scale, in the sense of the mean of squared error (theorem 3.2).

5 GPR from latent index variable

5.1 EM, gradient-based, and stochastic estimation methods

In this section, we will present various means to estimate statistical models, restricting ourselves to the typical case where all the required probabilities have a density with respect to Lebesgue-measure. We further apply these to previously presented Factor analysis, and later on to GPR. Let us start with a well-known Expectation Maximization (EM) algorithm, originally popularized by Dempster, Laird & Rubin, (1977). This method can be used to find the maximas of likelihood function (or posterior probability density function), in the case that part of the data is unobserved. We immediately recognize that in Factor analysis, the true scores U are not directly observed, but only the answers to questions, Q. Hence, we have an "incomplete data" situation, where only part of the data (U, Q) is observed. Instead of the original work, we introduce the method following less heavy notation of Cappé et al., (2005). That is, we assume (U, Q) have a product measure $\mu \otimes \lambda$ and densities $\{f(u, q; \theta)\}_{\theta \in \Theta}$ with respect to it. θ are model parameters, and we investigate the task of maximizing a function $\theta \mapsto f(u; \theta) := f(u, q; \theta)$, where dependence on the observed data q has been made implicit. This changes nothing, but saves us from the trouble of writing out the constant q (that is constant *after* it has been observed, when used in estimation). $L(\theta) = \int f(u; \theta) d\lambda(u)$ must exist, and it stands for the likelihood of data when unobserved factor u is marginalized away. Thus,

$$p(u; \theta) := f(u; \theta)/L(\theta) \tag{5.1}$$

is the conditional probability density function of the unobserved u, given q.

Definition 5.1 (Intermediate Quantity of EM)**.** *The* intermediate quantity *of EM is the family* $\{\mathcal{Q}(\cdot, \theta')\}_{\theta' \in \Theta}$ *of real-valued functions on* Θ*, indexed by* θ'*, and defined by*

$$\mathcal{Q}(\theta, \theta') = \int \log f(u; \theta) p(u; \theta') du. \tag{5.2}$$

EM-algorithm proceeds by iteratively alternating between computing the intermediate quantity (the integral) and maximizing, or increasing, the resulting expression with respect to θ, which is then set as the θ' for the next iteration. For EM-algorithm to be useful, one must be able to compute the integral in

closed form. In many interesting situations (as in Factor analysis) this can be done. Next we will state a proposition which gives the reader an intuition for why EM-algorithm works, and we will prove it after the actual algorithm is provided. Proofs on the convergence of the algorithm are just briefly discussed, but can be found for example in Cappé et al., (2005). Originally, they were provided by Wu, (1983). Recall, that for the *log-likelihood*

$$\ell(\theta) := \log L(\theta), \tag{5.3}$$

maximum of $\ell(\theta)$ equals that of $L(\theta)$, because logarithm is a monotone and bijective function. Now.

Proposition 5.1. *Assume that*
(i) The parameter set Θ is an open subset of $\mathbb{R}^d$ for some integer d.
(ii) For any $\theta \in \Theta$, $L(\theta)$ is positive and finite.
(iii) For any $(\theta, \theta') \in \Theta \times \Theta$, $\int \mid \bigtriangledown_\theta \, p(u;\theta)|p(u;\theta')du < \infty$, where $\bigtriangledown_\theta p$ is the gradient vector of p with respect to θ.
Then, for any $(\theta, \theta') \in \Theta \times \Theta$,

$$\ell(\theta) - \ell(\theta') \geq \mathcal{Q}(\theta, \theta') - \mathcal{Q}(\theta', \theta'), \tag{5.4}$$

where the inequality is strict unless $p(\cdot;\theta)$ and $p(\cdot;\theta')$ are equal almost everywhere (that is, with respect to Lebesgue-measure).

Before the proof, notice that assumption (*iii*) implies that all of the members of family of distributions $\{p(\cdot;\theta)\}_{\theta\in\Theta}$ are absolutely continuous with respect to other members. That is, sets with null probability are the same across entire family. Thus, the *Kullback-Leibler divergence* (*relative entropy*),

$$\mathcal{D}_{KL}(p(\cdot;\theta), p(\cdot;\theta')) := -\int p(u;\theta) \log \frac{p(u;\theta')}{p(u;\theta)} du, \tag{5.5}$$

is well-defined for all members of the family, with the conventions $0 \log 0 = 0$ and $0 \log \infty = 0$. Proposition itself implies that whenever we increase the $\mathcal{Q}(\theta, \theta')$ from the baseline of the $\mathcal{Q}(\theta', \theta')$, the corresponding increase in $\ell(\theta)$ over $\ell(\theta')$ is at least as large. With the following algorithm, this gives us an ascending sequence, $\{\ell(\theta^{(1)}), \ell(\theta^{(2)}), \ldots\}$, toward the maximum obtainable value.

Algorithm 5.1. *Make an initial guess, $\theta^{(0)}$, for the optimal value of θ, and iterate*
E-step: Determine the function $\theta \mapsto \mathcal{Q}(\theta; \theta^{(i)})$;
M-step: Choose $\theta^{(i+1)}$ to be any value $\theta \in \Theta$ that maximizes $\mathcal{Q}(\theta; \theta^{(i)})$.

Iterating this ascent algorithm may lead to local maxima, instead of global, if such exist. Hence, a good choice of the starting value, near the global maximum, helps in avoiding local ones. Furthermore, additional conditions are needed to ensure that the algorithm indeed ever converges to a value. These are (Cappé et al., 2005, chap. 10): Any level set $\{\theta \in \Theta : \ell(\theta) \geq \ell(\theta')\}$ is compact

and in interior of Θ, $\theta' \mapsto \int \log(p(u;\theta))p(u;\theta')du$ is continuous, and $\theta \mapsto L(\theta)$ and $\theta \mapsto \int \log(p(u;\theta))p(u;\theta')du$ are continuously differentiable on Θ. Here we only consider cases that satisfy the conditions of convergence. Let us now prove 5.1.

Proof of the proposition 5.1. From the equality 5.1, we see that

$$\begin{aligned}\mathcal{Q}(\theta,\theta') &= \int \log\left(L(\theta)p(u;\theta)\right)p(u;\theta')du \\ &= \ell(\theta) + \int \log(p(u;\theta))p(u;\theta')du,\end{aligned}$$

since $\ell(\theta) = \log L(\theta)$, $L(\theta)$ is constant with respect to u, and $p(u;\theta')$ is a probability density that integrates to 1. Same holds for $\mathcal{Q}(\theta',\theta')$, but with $\theta = \theta'$, of course. Now, observe that

$$\begin{aligned}&\mathcal{Q}(\theta,\theta') - \mathcal{Q}(\theta',\theta') \\ &\quad= \ell(\theta) - \ell(\theta') + \int \log(\frac{p(u;\theta)}{p(u;\theta')})p(u;\theta')du \\ &\quad= \ell(\theta) - \ell(\theta') - \mathcal{D}_{KL}(p(\cdot;\theta), p(\cdot;\theta')). \qquad (5.6)\end{aligned}$$

Since, by assumption (iii), $\mathcal{D}_{KL}$ is well-defined, it is only left to show that this Kullback-Leibler divergence is always positive and zero only when $p(\cdot;\theta) = p(\cdot;\theta')$. Consider general probability densities, g and p that are absolute continuous with respect to each other. Elementary inequality states that $\log x \leq x-1$, with equality only when $x = 1$. This is a consequence of concavity of the logarithm-function (function values lie below its tangents). Now,

$$\begin{aligned}(-1)\mathcal{D}_{KL}(g,p) &= \int \log(\frac{g(u)}{p(u)})p(u)du \\ &\leq \int (\frac{g(u)}{p(u)} - 1)p(u)du \\ &= \int (g(u) - p(u))du = 1 - 1 = 0.\end{aligned}$$

We have equality only if $\log(\frac{g(u)}{p(u)}) = \frac{g(u)}{p(u)} - 1$ for almost all u, which happens only if $g = p$ almost everywhere.

□

Benefits of EM are its independence regarding parameterization (invertible maps on θ), and the fact that possible parameter constraints are automatically incorporated. As evident, EM-algorithm is useful mainly when one can easily compute the expectation in Intermediate quantity *and* solve the maximum of this quantity in closed form. When EM is not feasible, gradient-based methods may offer an alternative (Cappé et al., 2005, chap. 10). With gradient-based approach, one does not need to be able to find a closed form solution to maximum. Also, faster convergence may be achieved via second order information,

as discussed below. For the use of gradient-based methods with partially observed data, it helps if the *Fisher's and Louis' identities* hold. Let us denote the Hessian matrix of second derivatives with $\bigtriangledown^2_\theta f := [\frac{\partial^2}{\partial\theta_i \partial\theta_j} f]_{ij}$. By noticing that derivative is just a limit of approximating difference quotient, we may often use Lebesgue's Dominated convergence theorem, and Mean-value theorem of differential calculus, to justify the change in order of integration and differentiation (Klenke, 2008, Theorem 6.28). In these cases, the following proposition can be proved.

Proposition 5.2 (Fisher's and Louis' identities)**.** *Assume that, in addition to assumptions of proposition 5.1, the following holds:*
a) $\theta \mapsto L(\theta)$ *is twice continuously differentiable on* Θ.
b) For any $\theta' \in \Theta$, $\theta \mapsto \int \log p(u;\theta) p(u;\theta') du$ *is twice continuously differentiable on* Θ. *In addition, for any* $k = 1, 2$ *and for any* $(\theta, \theta') \in \Theta \times \Theta$,

$$\int | \bigtriangledown^k_\theta \log p(u;\theta)| p(u;\theta') du < \infty,$$

and

$$\bigtriangledown^k_\theta \int \log p(u;\theta) p(u;\theta') du = \int \bigtriangledown^k_\theta \log p(u;\theta) p(u;\theta') du.$$

Then the Fisher's identity

$$\bigtriangledown_\theta \, \ell(\theta') = \int \bigtriangledown_\theta \log f(u;\theta)\Big|_{\theta=\theta'} p(u;\theta') du, \tag{5.7}$$

and the Louis' identity

$$\begin{aligned} -\bigtriangledown^2_\theta \, \ell(\theta') = &-\int \bigtriangledown^2_\theta \log f(u;\theta)\Big|_{\theta=\theta'} p(u;\theta') du \\ &+ \int \bigtriangledown^2_\theta \log p(u;\theta)\Big|_{\theta=\theta'} p(u;\theta') du \end{aligned}$$

holds. The second equality may be rewritten in the equivalent form

$$\begin{aligned} \bigtriangledown^2_\theta \ell(\theta') + \bigtriangledown_\theta \ell(\theta') \bigtriangledown_\theta \, \ell(\theta')^T = \int \Big[& \bigtriangledown^2_\theta \log f(u;\theta)\Big|_{\theta=\theta'} \\ & + (\bigtriangledown^2_\theta \log f(u;\theta)\Big|_{\theta=\theta'})(\bigtriangledown^2_\theta \log f(u;\theta)\Big|_{\theta=\theta'})^T \Big] p(u;\theta') du. \end{aligned} \tag{5.8}$$

Proof. (see proposition 10.1.6 in Cappé et al., 2005).

□

From Fisher's and Louis' identities we see that it is possible to compute the gradient vector and Hessian matrix at the point θ', even though we have not observed u (but do have a distribution for it, given θ'). That is, if we are able to compute the relevant expectations under $p(u;\theta')$, we can use gradient ascent methods to maximize the log-likelihood, instead of closed form solution.

Recall, that gradient, $\bigtriangledown_\theta \ell(\theta)$, points to the direction in Θ that gives fastest growth in $\ell(\theta)$. Hence, we can maximize $\ell(\theta)$ by iteratively taking

$$\theta^{(n+1)} = \theta^{(n)} + \gamma_n \bigtriangledown_\theta \ell(\theta^{(n)}),$$

where the scalar $\gamma_n \geq 0$ needs to adjusted in every iteration to ensure, at least, that the sequence $\{\ell(\theta^{(n)})\}$ is non-decreasing. This gives the *steepest ascent* algorithm. Here, in contrast to EM, we have an additional dilemma of choosing γ_n. If we choose this value badly, it may happen, for example, that we "jump" over the maximum, and decrease the likelihood function. A good choice is to solve, in each iteration, a simpler optimization problem, called a *line-search*:

$$\gamma_n = \arg\max_{\gamma \geq 0} \ell\left(\theta^{(n)} + \gamma \bigtriangledown_\theta \ell(\theta^{(n)}\right).$$

Steepest ascent algorithm has rather similar converge properties as the EM-algorithm (Cappé et al., 2005). Faster gradient ascent algorithm, with quadratic convergence properties, is achieved by setting the second order Taylor approximation,

$$\ell(\theta) \approx \ell(\theta') + \bigtriangledown \ell(\theta')(\theta - \theta') + \frac{1}{2}(\theta - \theta')^T \bigtriangledown^2 \ell(\theta')(\theta - \theta'),$$

to zero. As in the steepest ascent algorithm, we want to guard against badly chosen step sizes and, instead of simple solution, iterate

$$\theta^{(n+1)} = \theta^{(n)} + \gamma_n H^{-1}(\theta^{(n)}) \bigtriangledown_\theta \ell(\theta^{(n)}), \tag{5.9}$$

where γ_n is chosen from the line-search and $H(\theta^{(n)})$ is either the Hessian matrix $\bigtriangledown^2_\theta \ell(\theta^{(n)})$, or some other convenient approximating matrix (Cappé et al., 2005). When $\gamma_n \equiv 1$ and H is the Hessian, this algorithm is called *Newton-Raphson algorithm*. Via Louis' identity, we can find the Hessian in missing/latent data case.

Both, gradient ascent and EM approaches, require that one is able to compute, slightly different, integral with respect to $p(u; \theta')$ in closed form. Natural next step is to ask whether one can loosen this requirement, extending to case where the integral is not easily solvable. Very traditional remedy to intractable integrals is to use approximating Monte Carlo integration. If we can simulate a sequence of i.i.d. (independent and identically distributed) random values, $(U_i)_{i=1}^n$, from a probability distribution P, then the Strong law of large numbers implies that for a Borel-measurable function g,

$$\frac{1}{n}\sum_{i=1}^{n} g(U_i) \overset{a.s.}{\to} E[g(U_1)] = \int g(u) dP(u),$$

where a.s. refers to almost sure convergence (convergence with probability 1). Also, finite first and second moments are required from U_1. If we can simulate random values from the distribution with density $p(\cdot; \theta')$, we immediately get

a stochastic version of the EM-algorithm, where the intermediate quantity is computed from m simulated values as

$$\mathcal{Q}_m(\theta, \theta') := \frac{1}{m} \sum_{j=1}^{m} \log f(u^{(j)}; \theta) \approx \mathcal{Q}(\theta, \theta'). \tag{5.10}$$

Natural name for resulting algorithm is the Monte Carlo Expectation Maximization (MCEM) algorithm. This algorithm is found to perform better when number of simulated values is increased in each iteration of the algorithm (Cappé et al., 2005). As EM-iteration proceeds, its step sizes tend to get smaller, also requiring for smaller Monte Carlo estimation error.

Similarly, we can compute an approximation

$$\bigtriangledown \ell(\theta') \approx \frac{1}{m} \sum_{j=1}^{m} \bigtriangledown_\theta \log f(u^{(j)}; \theta)\Big|_{\theta=\theta'}, \tag{5.11}$$

and proceed with gradiant-based approaches (same natural extension applies to Hessian via Louis' identity). It should be obvious that Monte Carlo approach introduces additional estimation error to EM-algorithm, and is a lot more computationally heavy procedure. Still, it may be a sufficient work-around when one is unable to otherwise solve the required integral.

For the gradient-based approach, Monte Carlo approximation is very close to classical *Stochastic gradient algorithm*. Convergence of this type of algorithm was originally shown by Robbins & Monro, (1951). Here, one defines a deterministic sequence $(\gamma_n)_{n=1}^{\infty}$, with the properties

$$\gamma_n > 0, \qquad \lim_{n\to\infty} \gamma_n = 0, \qquad \sum_n \gamma_n = \infty.$$

If now random variable Y_n is a noisy observation of a real-valued function h at $\theta^{(n-1)}$, $h(\theta^{(n-1)})$, then the sequence

$$\theta_n = \theta_{n-1} + \gamma_n Y_n \tag{5.12}$$

can be shown to converge in probability to the root of the equation h. If we set $Y_n = \bigtriangledown_\theta \log f(u^{(n)}; \theta^{(n-1)})$, where $u^{(n)}$ is a sample from the density $p(\cdot, \theta^{(n-1)})$, we have ended up with Robbins-Monro framework. Now, we have a collection of σ-algebras $\{\mathcal{F}_n\}$, known as the *filtration*, and defined by

$$\mathcal{F}_n = \sigma(\theta^{(0)}, u^{(0)}, u^{(1)}, \ldots, u^{(n)}).$$

Now $u^{(n)}|\mathcal{F}_{(n-1)} \sim p(\cdot|\theta^{(n-1)})$, and thus, because of Fisher's identity, $E[Y_n|\mathcal{F}_{(n-1)}] = \bigtriangledown_\theta \ell(\theta^{(n-1)})$. Hence, we can interpret Y_n as noisy observation of $\bigtriangledown_\theta \ell(\theta^{(n-1)}) =: h(\theta_{n-1})$, for which we are seeking root for. That is, according to properties of expectation, we may write $Y^{(n)} = \bigtriangledown_\theta \ell(\theta^{(n-1)}) + \xi_n$, with

$$\xi_n = \bigtriangledown_\theta \log f(u^{(n)}; \theta^{(n-1)}) - E[\bigtriangledown_\theta \log f(u^{(n)}; \theta^{(n-1)})|\mathcal{F}_{n-1}].$$

Now, $\{\xi_n\}$ is a (martingale difference) noise sequence with $E[\xi_n] = 0$ for all n.

While gradient-based methods also take steps in the direction of the gradient, Robbins-Monro approach does not require a line-search. When value of the log-likelihood function is computationally expensive, or entirely unfeasible, to evaluate, then stochastic gradient methods may offer a good alternative. In practice, convergence may be slow if sequence $\{\gamma_n\}$ is chosen badly. Various modifications to algorithm exist (Cappé et al., 2005). For example, one may take, in each iteration, instead of one sample $u^{(n)}$, several samples $\{u^{(n,j)}\}$. Then a Monte Carlo estimate may be used as $Y_n = \frac{1}{m}\sum_{j=1}^{m} \bigtriangledown_\theta \log f(u^{(n,j)}; \theta^{(n-1)})$. While this should reduce noise from the sequence, there appears to be little theoretical guidance for the choice of sample-size m.

Above reviewed methods can be used to find a point estimate for θ, which is the Maximum likelihood estimate. There exist more sophisticated methods for estimating entire distribution for θ, that is, Bayesian a posteriori distribution (see eq. 3.12). These might, for example, set up a Markov Chain sequence that, after some iterations, is known to produce random values from the correct (invariant) distribution. From these values, one may then compute any desired Monte Carlo estimate, such as a posteriori mean or variance of the parameter θ. Also, unknown U can be estimated *as a parameter* in this framework. An umbrella term for methods with this strategy is Markov chain Monte Carlo (McMC) method. One such method is the Gibbs sampler, in which one chooses an arbitrary initial value $\theta^{(0)}$ and iterates

Algorithm 5.2 (Gibbs sampler)**.** *Update the current state* $\theta^{(k)} = (\theta_1^{(k)}, \ldots, \theta_m^{(k)})$ *to a new state* $\theta^{(k+1)}$ *as follows. For* $i = 1, \ldots, m$*: Simulate* $\theta_i^{(k+1)}$ *from the full conditional distribution of* θ_i *given other components, that is, from*

$$P(\cdot|\theta_1^{(k+1)}, \ldots, \theta_{i-1}^{(k+1)}, \theta_{i+1}^{(k)}, \ldots, \theta_m^{(k)}).$$

After each component has been updated, initiate new iteration.

After some amount of iterations, this algorithm starts producing values from the *joint* distribution of θ, or that of $(\theta|D)$, if we condition for some observed data D (Geman & Geman, 1984; Cappé et al., 2005). Due to properties of Gaussian distribution, full conditional distributions of Factor analysis model can be derived in closed form. Thus, this approach can be also applied to estimation of factor analysis, and unobserved value u in it (Lopes & West, 2004). Unfortunately, in depth discussion of McMC methods is out of the scope here. Now, after this, unsatisfyingly short, review of estimation methods, we will begin their application to problem at hand, that is, to GPR with latent index-observations.

5.2 Estimation of the model

As discussed above, we do not observe the true values u that index our Gaussian field $\{Y_u\}$, which with we wish to model the outcome variable Y. However, we

do have a model for the items q given this latent u, which is the Factor analysis model. For simplicity, we consider the case where latent covariance is identity matrix, $\Phi = I$. If we are not specifically interested about this latent covariance, the model is actually the same as in the general case. To see this, note that covariance matrix has (non-unique) Cholesky decomposition, $\Phi = MM^T$. Now, covariance of the observations is

$$\begin{aligned} Cov(Q) &= \Lambda_0 \Phi \Lambda_0^T + \Psi \\ &= \Lambda_0 M (\Lambda_0 M)^T + \Psi \\ &=: \Lambda_1 \Lambda_1^T + \Psi. \end{aligned}$$

That is, if Λ is estimated as free parameter(s), the density given by the model does not change depending on Φ. The subject matter *interpretation* may change considerably. However, when little is known a priori, assumption of independent latent sources is in line with a scientific principle of parsimony (Occam's razor), stating that simplest model should be chosen until proven wrong.

Let us now model $u \sim N(0, I)$ and $q \sim N(0, \Lambda\Lambda^T + \Psi)$ with Factor analysis, and y with a Gaussian field having the Squared exponential covariance structure (implying parameters v, W and σ_ξ, as previously set). In previous section it was shown that, in latent variable situation, it is possible to utilize the EM-algorithm with an intermediate quantity

$$\mathcal{Q}(\theta; \theta') = \int \log(f(y, q|u; \theta)) p(u|y, q; \theta') du, \tag{5.13}$$

where now $\theta = (W, v, \sigma_\xi, \Lambda, \Psi)$ contains all the unknown parameters in the likelihood of *joint model* $f(y, q|u, \theta) = \varphi_{(Y|U)}(y|u; \theta) f_{(Q|U)}(q|u; \theta)$. Factorization of the model follows from the fact that outcome Y is independent of items/questions Q given latent variable U. We further notice that distribution of the latent variable is independent of outcome variable and GPR-model predicting it, given items q. That is, $p(u|y, q; \theta) = p(u|q; (\Lambda, \Psi))$. Taking these into account, we find that

$$\begin{aligned} \mathcal{Q}(\theta; \theta') &= \int \log \varphi_{(Y|U)}(y|u; (W, v, \sigma_\xi)) p(u|q; (\Lambda', \Psi')) du \\ &\quad + \int \log f_{(Q|U)}(q|u; (\Lambda, \Psi)) p(u|q; (\Lambda', \Psi')) du \\ &=: \mathcal{Q}_\varphi(\theta, \theta') + \mathcal{Q}_f(\theta, \theta'). \end{aligned} \tag{5.14}$$

From above decomposition one immediately sees that maximum of $\mathcal{Q}_f$ is entirely independently of the GPR-model and outcome variable Y. It may thus be wise to estimate the simpler Factor analysis model $f_{(Q|U)}$ first, and set its maximum likelihood parameters before going to full estimation with respect to all parameters $(W, v, \sigma_\xi, \Lambda, \Psi)$. Almost all estimation algorithms converge faster and more reliably if one is able to start them near the maxima. Before doing this, let us review some helpful notations and concepts on matrix differentials.

In principle, likelihood function is a real-valued function of vector-valued parameter. However, in derivation, it is helpful to consider functions from scalar

to matrix. Thus, we introduce a derivative of some matrix Y with respect to scalar x as $\frac{\partial Y}{\partial x}$, where each element of matrix $[\frac{\partial Y}{\partial x}]_{ij}$ is a derivative of function $x \mapsto Y_{ij}(x)$ mapping x to $(i,j)^{\text{th}}$ element of the matrix Y. This provides a natural way to use the chain rule of differential calculus without loosing the matrix interpretation. Consider then a real-valued function from matrices, $X \mapsto f(X) \in \mathbb{R}$. Here, we set the derivative as matrix of equal size, with each element being ordinary derivative with respect to corresponding matrix element, $\frac{\partial f(X)}{\partial X} := [\frac{\partial f(X)}{\partial X_{ij}}]_{ij}$. Even though we arrange derivatives to a matrix, this approach is mostly the same as the gradient vector in ordinary vector calculus. As the trace is just a sum of the diagonal elements of the matrix, $Tr(Y) = \sum_{i=1}^{d} Y_{ii}$, linearity of the differential operator implies that $\frac{\partial \mathrm{Tr}(Y)}{\partial x} = \mathrm{Tr}(\frac{\partial Y}{\partial x})$. Furthermore, it is easy to see that $\mathrm{Tr}(X^T Y) = \mathrm{Tr}(XY^T) = \sum_{ij} X_{ij} Y_{ij} = \mathrm{vec}(X)^T \mathrm{vec}(Y)$, where $\mathrm{vec}(X)$ is a column vector of the elements of matrix X. That is, trace of the matrix product equals inner-product of their elements. Thus, we get an interpretation that a differential of real-valued matrix function $f(X)$ to direction H is

$$d_H(f(X)) = \mathrm{vec}(\frac{\partial f(X)}{\partial X})^T \mathrm{vec}(H) = \mathrm{Tr}(\frac{\partial f(X)}{\partial X}^T H),$$

which is just a differential of analogous vector function $\tilde{f}(\mathrm{vec}(X)) = \mathrm{vec}(f(X))$. Naturally, all operations must be defined. In fact, since inner-product defines a norm, we see that matrix-space $\mathbb{R}^{d \times m}$ is isomorphic to vector space $\mathbb{R}^{dm}$ with a canonical isomorphism, $\mathrm{vec} : \mathbb{R}^{d \times m} \to \mathbb{R}^{dm}$, that stacks the columns of matrix to a vector. Here, the matrix-space is endowed with *Frobenius norm*, $\| X \|_F = \sqrt{\sum_{i=1}^{d} \sum_{j=1}^{m} |X_{ij}|^2}$, and the vector space is endowed with a standard Euclidian norm. Thus, the adopted notation offers no greater generality than the familiar vector calculus, but sometimes it makes book-keeping a lot easier by allowing matrix formulas. Let us begin by reviewing few results regarding multidimensional differentials from this perspective. Specifically, we will need the derivative of the determinant function and inverse of the matrix.

Lemma 5.1. *If each element of the d-by-d invertible matrix Y is a differentiable function of the scalar x, then*

$$\frac{\partial det(Y)}{\partial x} = det(Y)\, Tr(Y^{-1} \frac{\partial Y}{\partial x}), \tag{5.15}$$

Furthermore, we have

$$\frac{\partial Y^{-1}}{\partial x} = -Y^{-1} \frac{\partial Y}{\partial x} Y^{-1}. \tag{5.16}$$

Proof. (Magnus & Neudecker, 1991, chapter 8, Theorems 2 and 3).

□

With these tools, let us now consider estimation of Factor analysis model via EM.

5.2.1 Estimation of Factor analysis model via EM-algorithm

We first derive a nice lemma for optimization in the context of Gaussian distributions, with which we get results directly for matrices, without element-wise arguments.

Lemma 5.2. *Let Q, Z and Y be square symmetric invertible matrices, and X and W arbitrary matrices. Assume that the dimensions of these matrices are such that all operations below make sense. Then,*

$$\frac{\partial(-\log|Q| - Tr(Q^{-1}Z))}{\partial Q}\Big|_{Q=Q^\star} = 0 \Rightarrow Q^\star = Z \tag{5.17}$$

$$\frac{\partial(\,Tr\left(Q^{-1}(WX^T + XW^T - XYX^T)\right)}{\partial X}\Big|_{X=X^\star} = 0 \Rightarrow X^\star = WY^{-1} \tag{5.18}$$

Proof. Using ordinary differentiation rules and the previous lemma on matrix derivatives, for any matrix-element ij following holds:

$$\begin{aligned}
&\frac{\partial(-\log|Q| - \mathrm{Tr}(Q^{-1}Z))}{\partial Q_{ij}} \\
&= -\mathrm{Tr}(Q^{-1}\frac{\partial Q}{\partial Q_{ij}}) - \mathrm{Tr}(Q^{-1}\frac{\partial Z}{\partial Q_{ij}} + \frac{\partial Q^{-1}}{\partial Q_{ij}}Z) \\
&= -\mathrm{Tr}(Q^{-1}\frac{\partial Q}{\partial Q_{ij}}) - \mathrm{Tr}(\frac{\partial Q^{-1}}{\partial Q_{ij}}Z) \\
&= \mathrm{Tr}(-Q^{-1}\frac{\partial Q}{\partial Q_{ij}}) + \mathrm{Tr}(Q^{-1}\frac{\partial Q}{\partial Q_{ij}}Q^{-1}Z) \\
&= \mathrm{Tr}(-Q^{-1}\frac{\partial Q}{\partial Q_{ij}}) + \mathrm{Tr}(Q^{-1}ZQ^{-1}\frac{\partial Q}{\partial Q_{ij}}) \\
&= \mathrm{Tr}\left((-Q^{-1} + Q^{-1}ZQ^{-1})\frac{\partial Q}{\partial Q_{ij}}\right) \\
&= \mathrm{vec}((-Q^{-1} + Q^{-1}ZQ^{-1})^T)^T[0,\ldots,0,\underbrace{1}_{(i+j)^{\mathrm{th}}},0,\ldots,0]^T.
\end{aligned}$$

Hence, the differential of left side of equation 5.17 is zero to all directions when $Q = Z$, meaning that, Z is the extremal point of the function $Q \mapsto -\log|Q| - \mathrm{Tr}(Q^{-1}Z)$. This proves the equation 5.17.

Other identity, 5.18, can be derived with similar application of differentia-

tion rules:

$$\begin{aligned}
&\frac{\partial(\mathrm{Tr}\left(Q^{-1}(WX^T + XW^T - XYX^T)\right)}{\partial X_{ij}} \\
&\quad = \mathrm{Tr}\Big(Q^{-1}\frac{\partial(WX^T + XW^T - XYX^T)}{\partial X_{ij}} \\
&\qquad + \frac{\partial Q^{-1}}{\partial X_{ij}}(WX^T + XW^T - XYX^T)\Big) \\
&\quad = \mathrm{Tr}\Big(Q^{-1}\Big(W(\frac{\partial X}{\partial X_{ij}})^T + \frac{\partial X}{\partial X_{ij}}W^T - \frac{\partial X}{\partial X_{ij}}YX^T - XY(\frac{\partial X}{\partial X_{ij}})^T\Big)\Big). \qquad (5.19)
\end{aligned}$$

Using the linearity of the Trace-operator, and the facts that in general $\mathrm{Tr}(A) = \mathrm{Tr}(A^T)$, $\mathrm{Tr}(A^TB) = \mathrm{Tr}(AB^T)$, $\mathrm{Tr}(AB) = \sum_{ij} A_{ij}B_{ji} = \sum_{ij} A_{ji}B_{ij} = \mathrm{Tr}(BA)$, and Q is symmetric, one sees that equation 5.19 takes the form

$$\begin{aligned}
&\mathrm{Tr}\left(Q^{-1}(2W^T - YX^T - Y^TX^T)\frac{\partial X}{\partial X_{ij}}\right) \\
&\quad = \mathrm{Tr}\left(Q^{-1}(2W^T - 2Y^TX^T)\frac{\partial X}{\partial X_{ij}}\right) \\
&\quad = \mathrm{vec}(Q^{-1}(2W^T - 2Y^TX^T))^T[0,\ldots,0,\underbrace{1}_{(i+j)^{\mathrm{th}}},0,\ldots,0]^T,
\end{aligned}$$

where the former equality comes from the assumption $Y^T = Y$. Now, $(2W^T - 2Y^TX^T) \equiv 0$ if we set $X = WY^{-1}$, showing extremal point of equation 5.18, and concluding the proof.

□

Let us now turn to closed form maximization of the intermediate quantity $\mathcal{Q}_f$, after which the EM-algorithm is completely defined for the Factor analysis model. Notice that trace of the scalar equals its value and verify that for each observation i

$$\begin{aligned}
&(q_i - \Lambda U_i)^T\Psi^{-1}(q_i - \Lambda U_i) \\
&\quad = \mathrm{Tr}\Big(\Psi^{-1}\Big(q_iq_i^T - q_i(\Lambda U_i)^T - (\Lambda U_i)q_i^T + \Lambda U_i(\Lambda U_i)^T\Big)\Big) \\
&\quad =: \mathrm{Tr}(\Psi^{-1}Z_i), \qquad (5.20)
\end{aligned}$$

where

$$Z_i = q_iq_i^T - q_iU_i^T\Lambda^T - \Lambda U_iq_i^T + \Lambda U_iU_i^T\Lambda^T.$$

When derivating with respect to Λ, the term $\Psi^{-1}q_iq_i^T$ vanishes as a constant, and derivative of $\mathrm{Tr}(\Psi^{-1}Z_i)$ can be recognized as of the form 5.18. From the Gaussian assumptions, log-likelihood of the Factor analysis model for n observations is

$$\log f(q|U;\theta) = \sum_{i=1}^{n}\left\{-d\log(2\pi) - \frac{1}{2}\log|\Psi| - \frac{1}{2}(q_i - \Lambda U_i)^T\Psi^{-1}(q_i - \Lambda U_i)\right\}.$$

Now, if neglecting the additive $(-\frac{n}{2}(2\pi))$ and multiplicative $(\frac{1}{2})$ constants, we find that

$$\begin{aligned}\mathcal{Q}_f(\theta,\theta') &= E[\log f(q|U;\theta)|Q=q,\theta'] \\ &\propto E[\frac{1}{n}\sum_{i=1}^{n}\Big(-\log|\Psi| - Tr(\Psi^{-1}Z_i)\Big)\,|q,\theta'] \\ &= n\left(-\log|\Psi| - Tr\Big(\Psi^{-1}(\frac{1}{n}\sum_{i=1}^{n}E[Z_i|q,\theta'])\Big)\right), \qquad (5.21)\end{aligned}$$

where the expectation within parentheses can be solved using proposition 4.1. Note that our Factor analysis model assumed $E[U] = 0$ and $E[UU^T] = I$. Hence,

$$E[U_i|q,\theta'] = \Lambda'^T(\Lambda'(\Lambda')^T + \Psi')^{-1}q_i$$

and

$$E[U_iU_i^T|q,\theta'] = I - \Lambda'^T(\Lambda'(\Lambda')^T + \Psi')^{-1}.$$

Furthermore,

$$\begin{aligned}&E[Z_i|q,\theta'] \\ &= q_iq_i^T - q_iE[U_i|q,\theta']^T\Lambda^T - \Lambda E[U_i|q,\theta']q_i^T + \Lambda E[U_iU_i^T|q,\theta']\Lambda^T. \qquad (5.22)\end{aligned}$$

Above calculation gives us one, and thus all, iterations of the EM-algorithm:

Proposition 5.3 (EM updata for Factor analysis)**.** *Given previous values $\theta' = (\Lambda', \Psi')$, maximize $\mathcal{Q}_f$ with respect to Ψ by setting*

$$\Psi_{new} = diag\left(\frac{1}{n}\sum_{i=1}^{n}E[Z_i|q,\theta']\right), \qquad (5.23)$$

where $E[Z_i|q,\theta']$ is given by 5.22 and notation diag(A) refers to a diagonal matrix with diagonal equal to that of A. Then, maximize $\mathcal{Q}_f$ with respect to factor loading matrix Λ by setting

$$\Lambda_{new} = \left(\sum_{i=1}^{n}E[U_i|q,\theta']q_i^T\right)\left(\sum_{i=1}^{n}E[U_iU_i^T|q,\theta']\right)^{-1}. \qquad (5.24)$$

Proof. For the update with respect to Ψ, apply lemma 5.2 directly to equation 5.21, taking into account that the off-diagonal values are fixed to zero by assumption. Notice then that,

$$\begin{aligned}\frac{\partial}{\partial\Lambda}\mathcal{Q}_f(\theta,\theta') = \frac{\partial}{\partial\Lambda}\mathrm{Tr}\Big(\Psi^{-1}\Big((\frac{1}{n}\sum_{i=1}^{n}q_iE[U_i|q,\theta']^T)\Lambda^T \\ +\Lambda(\frac{1}{n}\sum_{i=1}^{n}E[U_i|q,\theta']q_i^T) - \Lambda(\frac{1}{n}\sum_{i=1}^{n}E[U_iU_i^T|q,\theta'])\Lambda^T\Big)\Big),\end{aligned}$$

and apply lemma 5.2. Multivariate Gaussian belongs to so called *exponential* family of distributions, ensuring that these stationary points can be shown to be maximas. Furthermore, Ψ, by construction, will be proper positive definite covariance matrix. We leave these details without further attention.

□

While EM-algorithm is a successful optimization routine for Gaussian models, recall that in chapter 4 we gave arguments for estimation via McMC methods. Thus, it is recommended that one uses McMC estimation when dealing with sufficiently small models. However, when there are very many items, there will be abundance of free parameters. In these cases, simple closed form solution, as in EM or other traditional likelihood-based methods (such as in Lawley & Maxwell, 1971), may turn out as useful. McMC methods rely on setting up a chain of random values for each free parameter and latent value, being able to verify the convergence of each chain to a desired distribution, and on collecting many samples from each chain. The samples are then used to estimate the quantities of interest. If we are dealing with p items and k uncorrelated latent dimensions, Factor analysis model involves up to $p(k+1)$ free parameters, $k(k-1)/2$ of which can be bound (Lopes & West, 2004). Recall, p may be in order of hundreds, while k is typically less than 10.

We now turn to estimation of GPR-model. Even though we directly observe the indexing variable, we immediately see that general closed form solution cannot be obtained, unless we first specify the covariance function $K(\theta)$. Even if we specify the covariance function, such as the Squared exponential of chapter 3, we find that each element depends nonlinearly on index-data and parameters. Hence, we face a lot more complex problem than optimization of Linear Gaussian model, such as Factor analysis. Yet, rather general approach can be derived for the noise-free case. By noise-free, we mean that indexing variable $u = x$ is directly observed. Let us first consider estimation in this case, before going to more complicated latent data situation.

5.2.2 Estimation of GPR model in noise-free case

Let us use the Squared exponential covariance of chapter 3 as an example. We need some estimate for the W, v and σ_ξ, collectively referred as $\theta = (W, v, \sigma_\xi)$. Only after we have set some values for these parameters, can we perform prediction by conditioning to observed data. There are many approaches to such an estimation problem, but one of the most standards is the Maximum-likelihood (ML) approach. Here, we simply state that the most likely values, $(\hat{W}, \hat{v}, \hat{\sigma_\xi})$, are those for which the data density is maximized, that is

$$(\hat{W}, \hat{v}, \hat{\sigma_\xi}) = \arg\max_\theta \varphi((y)_{i=1}^n | (x)_{i=1}^n, \theta). \tag{5.25}$$

Many methods for function extremizing exist, implementations of which can be found from software for numerical computation, or as separate programs. As this seems standard practice in the field, we take the ML-approach of Rasmussen & Williams, (2006), where each element of θ is maximized with a conjugate gradient ascent method. Hence, one only needs the partial derivatives for each element of θ, and the (conjugate) gradient ascent routine seems to perform well in this simple case.

Writing the observed outcome data as vectors, $Y_{data} =: y \in \mathbb{R}^n$, our vector of observations follow the probability density

$$\varphi_{(Y|X)}(y|x;\theta) = (2\pi)^{-n/2}|K|^{-1/2}e^{-\frac{1}{2}y^T K^{-1}y},$$

where K is now a function of observations $(x_1, \ldots, x_n)$ and parameters $(W, v, \sigma_\xi) = \theta$. As previously shown, log-likelihood is then proportional to

$$\ell(\theta) = \log \varphi_{(Y|X)}(y|x;\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|K(\theta)| - \frac{1}{2}y^T K(\theta)^{-1}y, \tag{5.26}$$

where K is given, as in chapter three, by $\Sigma + \sigma_\xi I_n$, where Σ is formed as in equation 3.6. Using above lemma 5.1 on matrix derivatives, we immediately get derivatives of log-likelihood for the general $K(\theta)$. These are

$$\frac{\partial \ell(\theta)}{\partial \theta_k} = \frac{1}{2}y^T K(\theta)^{-1}\frac{\partial K(\theta))}{\partial \theta_k}K(\theta)^{-1}y - \frac{1}{2}Tr(K(\theta)^{-1}\frac{\partial K(\theta)}{\partial \theta_k}), \tag{5.27}$$

where k indexes all the free parameters that are to be estimated. The computation so far generalizes to every differentiable covariance function. From this point onwards, we need to plug-in the specific choice.

Thus, it is left to compute $\frac{\partial K(\theta))}{\partial \theta_k}$. For the Squared exponential covariance, derivative with respect to σ_ξ is simply I_n, and that with respect to v equals Σ. However, elements of W, $\theta_k = W_{kk}$, are slightly more elaborate case. Let us express the covariance function (eq. 3.2) as $vg(W)$ where g is real-valued function $g(W) := \exp\left(-\frac{1}{2}(x_i - x_j)^T W^{-1}(x_i - x_j)\right)$. Again with traditional differentiation rules and above lemma, we find that

$$\frac{\partial g(W)}{\partial W_{kk}} = -\frac{1}{2}\alpha^T \frac{\partial W}{\partial W_{kk}}\alpha e^{-\frac{1}{2}(x_i - x_j)^T W^{-1}(x_i - x_j)}, \tag{5.28}$$

where $\alpha := W^{-1}(x_i - x_j)$. Thus, for each element (i, j) of $\frac{\partial K(\theta))}{\partial W_{kk}}$ we need to compute above values, evaluate $v\frac{\partial g(W)}{\partial W_{kk}}$, and further assign to equation 5.27 used in the chosen gradient ascent algorithm. However, before this, one small matter requires attention. In the definition of the Squared exponential, we constrained parameters to be positive. This applies to σ_ξ as well, since it does not make sense for variance to be negative. Standard trick for applying this constraint in gradient optimization is to optimize *logarithm* of parameters instead of the true parameters. That is, in gradient ascent, we take each parameter to be $\theta_i = \log \theta_i \star$. In result, all functions of theta, such as $g(W(\theta^\star))$ take a form $g(W(\exp(\theta)))$. As now the true $\theta^\star$ is exponential, it is constrained to be positive.

These, more or less standard, computations apply only when indexing variable $x = u$ has been directly observed. For us, the latent u of interest is not directly observed. Only some noisy estimate, $x = A^T q$, of it can be observed. Notice, that if we have estimated the Factor analysis model and use the proposition 4.1, we get an optimal (see theorem 3.2) estimate $x = A^t q = \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}q$, as well as error estimate for it. Notice however, that error estimate given by $E[UU^T|Q = q, \Lambda, \Psi]$ is an overly optimistic one, as it does not take into account the uncertainty in estimates for Λ and Ψ. Yet, it is comforting that *some* estimate can be easily derived.

5.2.3 Estimation of GPR model with noisy indices

We would like to do something similar to GPR model that we did for Factor analysis, that is, estimate it despite the fact that we have not observed predictor variable U. In the beginning of this chapter, we gave a general discussion about the methods for this kind of situation. Let us now use this information to examine the case of GPR model when index variable is a latent variable of Factor analysis model. In the case of EM-algorithm, we would need to be able to compute the quantity

$$\mathcal{Q}_\varphi(\theta;\theta') = \int \log \varphi(y|u;\theta) p(u|q;\theta') du, \tag{5.29}$$

where $p(u|q;\theta)$ is the Gaussian density given by proposition 4.1, and

$$\log \varphi(y|u;\theta) =: (\ell(\theta)|U=u)$$

is the log-likelihood given $U = u$. Unfortunately, now covariance structure in this likelihood, $K(\theta) =: K(\theta, (U_1, \ldots, U_n))$, depends on latent observations on a nonlinear way. This is likely to make the integral intractable. Hence, EM-approach, does not seem to be a good choice. Even if we would be able to solve the integral, we would still face the same problem as in the noise-free case. If $K(\theta)$ is complicated function, integral is not likely to be much simpler, and it will be difficult to optimize this in closed form. We would also need to be able to integrate over the absolute value of determinant of $K(\theta)$, and to solve the resulting expression. This strategy could be investigated for simple covariance structures, like the linear one $C(u_i, u_j) = u_i^T W u_j$, or $C(u_i, u_j) = (u_i - u_j)^T W (u_i - u_j)$. For general covariance, we take it as infeasible, and proceed to other choices.

We could attempt to set-up the McMC chain, with posterior distribution of θ as its invariant distribution. However, to derive a simple sampler, like Gibbs sampler, we would need to solve so called *full conditional distribution* to all parameters θ. In general case, this cannot be done. Again, maybe for some special cases. It might be possible to construct some other sampler (Cappé et al., 2005, chap. 6 & 7). While undoubtedly computationally heavy, this approach could be examined by a researcher well-versed with McMC methods. Instead of Markov chains, we observe that a Monte Carlo estimate of integral in 5.29, almost, brings us back to the noise-free case. As in noise-free case, we still cannot solve the closed form maxima for Monte Carlo approximation of 5.29, as would be required for EM. However, we sure can compute the derivatives

$$\frac{\partial}{\partial \theta_i} \ell(\theta) \approx \frac{1}{m} \sum_{j=1}^{m} \frac{\partial}{\partial \theta_i} \log \varphi(y|u^{(j)};\theta), \tag{5.30}$$

where $(u^{(1)}, \ldots, u^{(m)})$ is an i.i.d sample from the distribution $p(\cdot|q;\theta')$, given by proposition 4.1. Thus, we *are* able to perform gradient ascent, in spite of the latent index variable. Clearly, $\frac{\partial}{\partial \theta_i} \log \varphi(y|u^{(j)};\theta)$ is computed just as in the

noise-free case where $x = u^{(j)}$. Hence, we can do a Monte Carlo version of the standard Gradient ascent algorithms, but what is the cost of this flexibility?

If we wish to compute gradient at all points $(u^{(j)})_{j=1}^m$, we need to invert n-by-n matrix $K(\theta)$ at least m times (see eq. 5.27). Inverting $K(\theta)$ takes in order of n^3 basic computer operations (Rasmussen & Williams, 2006; Rue & Held, 2005), where n is the number of observations[1]. Hence, we need in order of mn^3 operations per each gradient evaluation. In practice, both m and n, may exceed 1000, meaning that every single iteration of gradient ascent takes over 10^{12} operations. Recall also that, for gradient ascent to be effective, it is desirable to perform line-searches in order to find correct step-size (to gradient direction). Hence, for each gradient-value, one should evaluate $\log \varphi(y|\cdot;\theta)$ with several values, and this also contains the inverse of $K(\theta)$. Combine these observations with the fact that it may take several iterations of gradient ascent to get near the maxima, and even modern computers start to edge toward their maximum capacity.

An average personal computer fairly rapidly maximizes (or at least finds good) parameters in the noise-free case, for over 1000 observations. In this context, data sets from several hundreds to few thousands observations start to be fairly large, often largest available. However, the need to do every gradient evaluation m times for the Monte Carlo approximation starts to tax more heavily than is be desirable. According to Wikipedia[2], 10^{12} floating point operations take about 22-23 minutes of processing time on a modern personal computer, about 1.34 seconds if one computes using graphics processing card, and cirka one millisecond on current supercomputers. While graphics card implementation seems to drop the computational burden near noise-free case, in practice these approaches are complicated by slow transfer of data between graphics card memory and computer memory. We assume that target group for the method has no easy access to supercomputers. Thus, it is desirable to do something for the amount of computations.

Given the above discussion, it seems that the best way to estimate GPR model, in this latent variable context, is to construct a good Stochastic gradient algorithm. This sets the size of Monte Carlo sample to 1 per each iteration. As discussed, one may in practice take larger amount of Monte Carlo samples, but the point is that this amount can be significantly smaller than would be required for brute force Monte Carlo. However, classical Robbins-Monro-type algorithms are typically found to be inadequate in practice, only converging to correct maxima when initiated nearby (Gu & Zhu, 2001). Algorithms that supplement classical approach with second-order (Newton-Rhapson-type) information have lead to good results in more challenging estimation tasks (Gu & Zhu, 2001; Cai, 2010). It is relatively straightforward to derive the Hessian of $\log \varphi(y|u^{(j)};\theta)$ by derivating equation 5.27 again with respect to θ_j, using

[1] If $K(\theta)$ is, or we can make it, sparse, this may reduce number of operations to $\mathcal{O}(n)$, but we assume such trick is not available. Matrix is said to be sparse, if most of its elements are zeros.

[2] http://en.wikipedia.org/wiki/Orders_of_magnitude_(computing)

lemma on matrix derivatives 5.1. Letting indices k (in 5.27) and j run through all parameters gives elements of the Hessian matrix. Louis' identity can be used to parse together the Hessian of log-likelihood. These algorithms can be shown to converge even if latent index is drawn from the Markov chain (Cai, 2010). Hence, currently best estimation option might be something like the following algorithm, adopted from Cai, (2010). We present it only for completeness, but do not explicitly use it.

Algorithm 5.3 (Stochastic gradient algorithm for latent variable GPR)**.** *Take a sequence* $(\gamma_j)_{j=1}^{\infty}$ *satisfying*

$$\gamma_j > 0, \qquad \lim_{j\to\infty} \gamma_j = 0, \qquad \sum_j \gamma_j = \infty.$$

Use EM or maximum likelihood procedures to get initial values for Factor analysis parameters Ψ, Λ*. Run Gibbs sampler, starting from these values, until it appears to produce values from the desired distribution. Set* Γ_0 *to a positive definite matrix with as many elements as square of number of elements in* θ*. Choose* $\theta^{(0)}$ *and iterate, until sufficient convergence, from* $j = 1$

- *Stochastic imputation: using current value of* $\theta^{(j)}$ *draw* m_j *sets of latent index values from the Factor analysis Gibbs sampler (Lopes & West, 2004).*
- *Stochastic approximation: compute approximation for the gradient*

$$s_{j+1} = \frac{1}{m_j} \sum_{k=1}^{m_j} \bigtriangledown_\theta \ell(\theta^{(j)}, u_k),$$

where $\{u_k\}$ *are previously drawn latent index values and* $\bigtriangledown_\theta \ell(\theta^{(j)}, u_k)$ *is gradient for the full model log-likelihood (i.e. including combination of GPR and Factor analysis models). Compute recursive approximation for complete data expected information/Hessian matrix*

$$\Gamma_{j+1} = \Gamma_j + \gamma_j \left(\frac{1}{m_j} \sum_{k=1}^{m_j} - \bigtriangledown_\theta^2 \ell(\theta^{(j)}, u_k) - \Gamma_j \right).$$

- *Robbins-Monro update: Set the new parameter estimate to*

$$\theta^{(j+1)} = \theta^{(j)} + \gamma_j(\Gamma_{j+1}^{-1} s_{j+1}).$$

m_j may be either fixed value, or change as a function of iteration number (Cai, 2010). Due to Newton-Rhapson element, this Robbins-Monro -type algorithm takes large initial steps toward correct maxima. It further averages noise *over iterations* with decreasing sequence $(\gamma_j)_{j=1}^{\infty}$, which is why it converges, even though each iteration of gradient involves noise. Choice of the sequence $(\gamma_j)_{j=1}^{\infty}$ influences to practical success of algorithm. One might start with a choice of harmonic sequence, $(\frac{1}{j})_{j=1}^{\infty}$.

Explicit construction of this algorithm, not to mention possible two-step and stopping criterion modifications (Gu & Zhu, 2001), takes more parameter tweaking and details than is sensible to present here. One may also encounter, more or less, unexpected difficulties when implementing such algorithm. This is because Newton-Rhapson-like procedures do not generalize as well as first-order gradient ascent methods, but may require tuning for specific covariance structure. Vulnerability of simple Newton-Rhapson procedure (without line-search) can be easily shown with below example. Consider optimization of simple one-dimensional deterministic function.

$$g(w) = e^{-\frac{1}{2}w^2} \tag{5.31}$$

In this case, the gradient is

$$g'(w) = -we^{-\frac{1}{2}w^2},$$

and the Hessian is given by

$$g''(w) = (w^2 - 1)e^{-\frac{1}{2}w^2}.$$

Combining these, we see that single Newton-Rhapson iteration is of the form

$$w_{n+1} = w_n - g''(w_n)^{-1}g'(w) = w_n + \frac{w_n}{w_n^2 - 1}. \tag{5.32}$$

If we start the iteration from $w_0 > 1$, we will, in each iteration, move *farther* from the maxima $w = 0$. Even if we were doing line-searches to avoid wrong directions, it is evident that above iteration is numerically unstable near $w = \pm 1$. Thus, finding good coefficient γ_n numerically may be hindered by tendency of Newton-Rhapson algorithm to produce dubious step-sizes in some regimes of parameter space. A cursory look to Squared exponential covariance function is enough to tell that above example may have practical implications as well.

In any case, we recommend starting with a search for algorithm of the above type. Averaging *with respect to* $\theta^{(j)}$-sequence may improve the performance of this algorithm (Cappé et al., 2005). Compared to otherwise computationally expensive procedures in GPR estimation, this requires no further resources. There exist large literature, with convergence proofs, on Stochastic gradient procedures, and one may freely explore this. Instead of going to questions of practical implementation, we will continue with a question: is it actually possible to *use* this model after we have managed to estimate it? That is, can we make a prediction on a new point u given that we do not have the (now latent) index-observations for which to condition. Recall, that entire utility of GPR approach lies in the closed form formula 3.3 for conditional distributions, which allows handy predictive equations of chapter 3. Have we done all these efforts just to find out that we cannot use the model we have estimated? In next, and last, chapter we will come up with an approximative approach, originally due to Girard, (2004), that allows flexible use of GPR model in spite of the noise in index-observations. For this, the Squared exponential (or linear) covariance function will turn out to be a helpful choice.

6 Prediction with noisy observations

Ones we have managed to estimate the parameters of the covariance function, we want to make predictions, inter- and extrapolation, as in chapter 3. That is, via the use of the theorem 3.3 on the conditionals of Gaussian distribution. Recall, this was the primary motivation in GPR, allowing for automatic non-linear smoothing and prediction. If we have estimated the Factor analysis model, then it is possible to use some estimate for the unobserved latent/true values, u. This should reduce noise from the indexing variable of GPR, as proper estimate contains less noise than simple average of some items. One may estimate u for individual i either as average of simulated values from the Gibbs sampler, or simply via proposition 4.1 as the mean of conditional distribution given the observations of items q_i and estimated parameters:

$$\hat{u}_i = \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}q_i, \tag{6.1}$$

where Ψ and Λ are estimated Factor analysis parameters. Then, $\hat{u}_i$ will be normally distributed with "error" variance

$$\Upsilon = I - \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}\Lambda. \tag{6.2}$$

That is, although latent variable estimates from Factor analysis reduce noise, as do simple averaging scales, we still do not observe the noise-free index we wish for. In fact, reliability indices of psychometric scales are thought to be acceptable if over 0.7, reflecting the acceptance within the community for measurements with nearly 1/3 of variance being random noise. In practice, reliabilities of standard scales vary between 0.6 and 0.95.

In this case one might turn to an approximate Gaussian process approach, pioneered by Girard, (2004) in the context of sequential predictions. Here we use another Gaussian process that accommodates its covariance function to the noise in observations, in a correct manner. We present an example for the popular Squared exponential covariance function, which is one of the few that allow an analytic derivation. It might, however, be the case that a researcher has no access to questions q_i, but only to some summary x_i which estimates u_i less optimally than would be the case for estimate 4.1 with parameters derived from a large population sample. Sometimes researchers also wish to use some standard scoring derived from q_i, instead of the more optimal prediction that could be achieved in above described manner. In chapter four we defined this scoring scheme as A^Tq_i for some pre-defined matrix A. Clearly, this variable has

different covariance/error properties. As discussed in the chapter four, *between individual* (population) covariance of this variable decomposes as

$$Cov(A^T Q) = A^T Cov(Q)A = A^T \Lambda \Lambda^T A + A^T \Psi A,$$

where the part $A^T \Psi A$ corresponds to Gaussian measurement noise. Thus, in the case of pre-defined scale,

$$\Upsilon := \text{diag}(A^T \Psi A)$$

could be taken as error variance estimate for all individuals i, where $\text{diag}(A^T \Psi A)$ refers to diagonal matrix with diagonal values equal to those of $A^T \Psi A$. Reason why we do not use entire matrix $A^T \Psi A$ is that we are not interested about correlation of errors in the *population*, but about their magnitude in used measures. Let us now see how these error variance estimates can be put to use.

6.1 Noise incorporating covariance function

Again, before derivation of the method itself, we introduce couple results that are needed. First, Law of total covariance, is a consequence of the definition of conditional expectation. Second, is simply re-expression for the product of Gaussian density functions.

Lemma 6.1 (Law of total covariance)**.** *Given a σ-algebra $\mathcal{G}$, covariance of two random variables, X and Y, decomposes as*

$$Cov[X, Y] = E[Cov[X, Y|\mathcal{G}]] + Cov[E[X|\mathcal{G}], E[Y|\mathcal{G}]]. \tag{6.3}$$

Notice that σ-algebra $\mathcal{G}$ could be generated by some third variable, say Z. Then $\mathcal{G} = \sigma(Z)$, and $E[X|\mathcal{G}] = E[X|Z]$ etc.

Proof of lemma 6.1. Proof uses the definition of conditional expectation, linearity of expectation, and some algebraic manipulation. Simply check that

$$\begin{aligned}
Cov[X, Y] &= E[XY] - E[X]E[Y] \\
&= E[E[XY|\mathcal{G}]] - E[E[X|\mathcal{G}]]E[E[Y|\mathcal{G}]] \\
&= E[Cov[X, Y|\mathcal{G}] + E[X|\mathcal{G}]E[Y|\mathcal{G}]] - E[E[X|\mathcal{G}]]E[E[Y|\mathcal{G}]] \\
&= E[Cov[X, Y|\mathcal{G}]] + E[E[X|\mathcal{G}]E[Y|\mathcal{G}]] - E[E[X|\mathcal{G}]]E[E[Y|\mathcal{G}]] \\
&= E[Cov[X, Y|\mathcal{G}]] + Cov[E[X|\mathcal{G}], E[Y|\mathcal{G}]]
\end{aligned}$$

□

Remember that, if inverse of the matrix A exists, then it holds for the determinant that $|A^{-1}| = 1/|A|$, and $|AB| = |A||B|$, when the product is defined. Furthermore, for invertible matrices, $A + B = A(B^{-1} + A^{-1})B$. In addition, following result holds for matrices.

Lemma 6.2 (Woodbury matrix identity)**.** *Given that the inverses exist and all matrix multiplications are defined,*

$$(A+UBV)^{-1} = A^{-1} - A^{-1}U(B^{-1}+VA^{-1}U)^{-1}VA^{-1} \tag{6.4}$$

Proof. By direct verification,

$$\begin{aligned}
&(A+UBV)\left[A^{-1} - A^{-1}U(B^{-1}+VA^{-1}U)^{-1}VA^{-1}\right] \\
&\quad = I + UBVA^{-1} - (U+UBVA^{-1}U)(B^{-1}+VA^{-1}U)^{-1}VA^{-1} \\
&\quad = I + UBVA^{-1} - UB(B^{-1}+VA^{-1}U)(B^{-1}+VA^{-1}U)^{-1}VA^{-1} \\
&\quad = I + UBVA^{-1} - UBVA^{-1} = I
\end{aligned}$$

□

Setting U and V to identity matrices in above lemma, we get a useful decomposition

$$(A+B)^{-1} = A^{-1} - A^{-1}(B^{-1}+A^{-1})^{-1}A^{-1} = B^{-1} - B^{-1}(A^{-1}+B^{-1})^{-1}B^{-1}.$$

By using above identities we are able to prove the final lemma needed to derive the actual method of interest.

Lemma 6.3 (Product of Gaussians)**.** *Let φ_a be the density of the Gaussian distribution $N_d(a,A)$, and φ_b that of the distribution $N_d(b,B)$. Then the product of φ_a and φ_b is*

$$\varphi_a(x)\varphi_b(x) = z\varphi_h(x), \tag{6.5}$$

where φ_h is the density of the distribution $N_d(h,H)$, where $h = H(A^{-1}a+B^{-1}b)$ and $H = (A^{-1}+B^{-1})^{-1}$. z is given by

$$z = \frac{1}{(2\pi)^{-d/2}|A+B|^{1/2}} e^{\frac{1}{2}(a-b)^T(A+B)^{-1}(a-b)}, \tag{6.6}$$

which can be recognized as either the density of a distributed as $N_d(b,A+B)$ or density of b distributed as $N_d(a,A+B)$.

Proof. Check that

$$\begin{aligned}
&\varphi_a(x)\varphi_b(x) = (2\pi)^{-d/2}|A|^{-1/2}(2\pi)^{-d/2}|B|^{-1/2}\times \\
&e^{-\frac{1}{2}(x-a)^TA^{-1}(x-a)-\frac{1}{2}(x-b)^TB^{-1}(x-b)} \\
&= (2\pi)^{-d/2}|A|^{-1/2}|H^1|^{1/2}|B|^{-1/2}(2\pi)^{-d/2}|H|^{-1/2}\times \\
&e^{-\frac{1}{2}(x^T(A^{-1}+B^{-1})x-x^T(A^{-1}a+B^{-1}b)-(A^{-1}a+B^{-1}b)^Tx+a^TA^{-1}a+b^TB^{-1}b} \\
&= (2\pi)^{-d/2}|A(A^{-1}+B^{-1})B|^{-1/2}(2\pi)^{-d/2}|H|^{-1/2}\times \\
&e^{-\frac{1}{2}(x^T(A^{-1}+B^{-1})x-x^TH^{-1}H(A^{-1}a+B^{-1}b)-(A^{-1}a+B^{-1}b)^THH^{-1}x+a^TA^{-1}a+b^TB^{-1}b)}
\end{aligned}$$

$= (2\pi)^{-d/2}|A+B|^{-1/2}(2\pi)^{-d/2}|H|^{-1/2}\times$
$e^{-\frac{1}{2}(x^TH^{-1}x - x^TH^{-1}h + h^TH^{-1}x + h^TH^{-1}h - h^TH^{-1}h + a^TA^{-1}a + b^TB^{-1}b)}$

$= \varphi_h(x)(2\pi)^{-d/2}|A+B|^{-1/2}e^{-\frac{1}{2}(a^TA^{-1}a + b^TB^{-1}b - h^TH^{-1}h)}.$

Thus, it is left to show that the term after $\varphi_h(x)$ equals z. To do this, we must show that $a^TA^{-1}a + b^TB^{-1}b - h^TH^{-1}h$ in the exponent equals $(a-b)^T(A+B)^{-1}(a-b)$. This can be done using Woodbury matrix identity to H, expanding with respect to both, A and B. An observation that $A(A+B)^{-1} = (I + A^{-1}B)^{-1} = (A^{-1}+B^{-1})^{-1}B^{-1}$ is also needed. Now, notice that

$$\begin{aligned}
h^TH^{-1}h &= (A^{-1}a + B^{-1}b)^TH(A^{-1}a + B^{-1}b) \\
&= a^TA^{-1}HA^{-1}a + b^TB^{-1}HB^{-1}b + 2a^TA^{-1}HB^{-1}b \\
&= a^TA^{-1}a - a^T(A+B)^{-1}a + b^TB^{-1}b - b^T(A+B)^{-1}b + 2a^TA^{-1}A(A+B)^{-1}b \\
&= a^TA^{-1}a + b^TB^{-1}b - \Big(a^T(A+B)^{-1}a - 2a^T(A+B)^{-1}b + b^T(A+B)^{-1}b\Big) \\
&= a^TA^{-1}a + b^TB^{-1}b - (a-b)^T(A+B)^{-1}(a-b).
\end{aligned}$$

Replacing $h^TH^{-1}h$ in $\varphi_h(x)(2\pi)^{-d/2}|A+B|^{-1/2}e^{-\frac{1}{2}(a^TA^{-1}a + b^TB^{-1}b - h^TH^{-1}h)}$ with above derived version completes the proof.

□

Next, we will start with estimated Factor analysis and GPR models, under Squared exponential covariance. We will derive the mean and covariance of (non-Gaussian) stochastic process indexed by noisy observations. For observations Y_i and Y_j,

$$E[Y_i|U_i] = 0, \tag{6.7}$$

and

$$Cov[Y_i, Y_j|U_i, U_j] = C(U_i, U_j). \tag{6.8}$$

Above we constructed a model, according to which, $X_i = U_i + \Xi_i$, where $\Xi_i \sim N_d(0, \Upsilon)$. Here, d is the number of scales, or dimension of the scale, depending on whether "scale" is used to refer to a multidimensional construct or a unidimensional one. There is no practical difference. Now, U_i is a vector of "true" values of behavioral dimensions/traits for the individual i. Ξ_i is measurement noise/error whose covariance is estimated from the Factor analysis model, as discussed in the above chapter. It turns out that we can still compute the mean and covariance of the resulting process, *as a function of the observed* X. At this point, we consider Factor analysis model only regarding to what it tells about the error Ξ. Due to symmetry of Gaussian distribution, $X = u + \Xi \sim N(u, \Upsilon)$ implies $U = x + \Xi \sim N(x, \Upsilon)$, provided that we start by thinking u as non-random. More principled way to derive this result would

be to think that we do not wish to say anything about the distribution of U *a priori*, and use *improper* Bayesian prior, 1. Then, using Bayes theorem, we see that posterior distribution of U given X is equal to $N(X, \Upsilon)$.

Let us assume in the following that the GPR-model of chapter three holds for the true value U. As now X is a measurable function (sum) of both, U and Ξ, it is $\sigma((U, \Xi))$-measurable. Since $E[Y|U] = 0$ by the model assumption, according to Tower property,

$$E[Y_{U_i}|X_i] = E[E[Y_{U_i}|U_i]|X_i] = 0. \tag{6.9}$$

Similarly, due to Law of total covariance,

$$\begin{aligned} &Cov[Y_{U_i}, Y_{U_j}|X_i, X_j] \\ &\quad = E[Cov[Y_{U_i}, Y_{U_j}|U_i, U_j]|X_i, X_j] + Cov[E[Y_{U_i}|U_i], E[Y_{U_j}|U_j]|X_i, X_j] \\ &\quad = E_{(U_i,U_j)}[Cov[Y_{U_i}, Y_{U_j}|U_i, U_j]|X_i, X_j]. \end{aligned} \tag{6.10}$$

Let us now denote the covariance function of noisy observations as

$$C_n(x_i, x_j) := Cov[Y_{U_i}, Y_{U_j}|X_i = x_i, X_j = x_j].$$

By noticing that individual/observation i is independent from j, from equation 6.10 we find that

$$C_n(x_i, x_j) = \int\int C(u_i, u_j)\varphi_{x_i}(u_i)\varphi_{x_j}(u_j)du_i du_j, \tag{6.11}$$

where φ_{x_i} is the density of $N_d(x_i, \Upsilon)$, and φ_{x_j} that of $N_d(x_j, \Upsilon)$.

Theorem 6.1 (Covariance function for noisy observations)**.** *Let u_i and u_j be observed with noise, such that $x_i = u_i + \xi_i$ and $x_j = u_j + \xi_j$, where ξ_i and ξ_j are distributed as $N(0, \Upsilon)$. Let $\{Y_u\}$ be a Gaussian process with zero mean and Squared exponential covariance function of equation 3.2. Then, the covariance of observations Y_i and Y_j, given the noisy observations x_i and x_j, is*

$$C_n(x_i, x_j) = v|I_d + 2W^{-1}\Upsilon|e^{-\frac{1}{2}(x_i - x_j)^T(W+2\Upsilon)^{-1}(x_i - x_j)} \tag{6.12}$$

Proof. Squared exponential covariance function $C : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ of equation 3.2 is of the form

$$C(u_i, u_j) = ve^{-\frac{1}{2}(u_i - u_j)^T W^{-1}(u_i - u_j)} =: \gamma\varphi_{u_j}(u_i),$$

where φ_{u_j} is the density of distribution $N_d(u_j, W)$, and

$$\gamma = v(2\pi)^{d/2}|W|^{1/2}.$$

According to equation 6.11, to find the covariance function $C_n(x_i, x_j)$, we need to evaluate the integral

$$C_n(x_i, x_j) = \gamma \int\int \varphi_{u_j}(u_i)\varphi_{x_i}(u_i)\varphi_{x_j}(u_j)du_i du_j. \tag{6.13}$$

Using the product of Gaussians equation (6.5), with analoguous indexing for the constant z as for Gaussian densities, and integrating over u_i, we find that

$$\begin{aligned}
&\int \varphi_{u_j}(u_i)\varphi_{x_i}(u_i)du_i \\
&\quad = z_{x_i}(u_j)\int \varphi_h(u_i)du_i \\
&\quad = z_{x_i}(u_j) \\
&\quad = (2\pi)^{-1/2}|W+\Upsilon|^{-1/2}e^{-\frac{1}{2}(u_j-x_i)^T(W+\Upsilon)^{-1}(u_j-x_i)}
\end{aligned}$$

Using the product of Gaussians equation again and further integrating this result with respect to u_j, we get

$$\begin{aligned}
\gamma^{-1}C_n(x_i,x_j) &= \int z_{x_i}(u_j)\varphi_{x_j}(u_j)du_j \\
&= (2\pi)^{-d/2}|W+2\Upsilon|^{-1/2}e^{-\frac{1}{2}(x_i-x_j)^T(W+2\Upsilon)^{-1}(x_i-x_j)}.
\end{aligned}$$

By multiplying both sides with a constant γ, we arrive to the desired result.

□

Thus, for noisy observations, this C_n is the covariance function of the observations. However, it is not the covariance function between Y_u, "observed" in the "true" value $u \in \mathbb{R}^d$, and Y_x, observed for the "noisy index" x_i. This covariance is also needed, because we are interested about the predictive means of the process for the true values. This is easily obtained from theorem 6.1 by letting $2\Upsilon =: (\Upsilon_u + \Upsilon_{x_i}) \to (\Upsilon_{x_i})$, reflecting the fact that uncertainty of another input is disappearing. Thus, we can let the covariance function depend on whether input was observed with noise or not.

Unfortunately, while deriving covariance for the noisy observations, we have ended up with a process that is *not* Gaussian. This implies that formulas allowing easily tractable analytic conditionals no longer exist (theorem 3.3 does not hold). Letting $C(U)$ denote covariance structure derived from C and true observations U, characteristic function takes the form

$$\begin{aligned}
E[e^{iY^T\theta}|X] &= E_U[E_{(Y|U)}[e^{iY^T\theta}]|X] \\
&= E_U[e^{-\frac{1}{2}\theta^T C(U)\theta}|X] \\
&= \int e^{-\frac{1}{2}\theta^T C(u)\theta}\varphi_{(u|X)}(u)du,
\end{aligned}$$

where C is highly non-linear function of U, involving another exponential function. Thus, the characteristic function cannot be of the Gaussian form $e^{-\frac{1}{2}\theta^T C_n(X)\theta}$, where C_n is matrix depending only on X.

While it is not easy to compute conditional distributions for this new process that is no more Gaussian, let us recall why we started with Gaussian processes at first place: They are flexible approximations to non-linear functions, provided that we find suitable covariance. As we now have derived a suitable covariance

in theorem 6.1, we can simply use this as a covariance for the approximating Gaussian process. As means are zero for both processes, this can also be seen as a *second order approximation.*

In practice, we proceed just as in chapter three, but replace C_n for covariance function C for actual observations, and

$$C_p(x_i, u_j) := v|I - W^{-1}\Upsilon|^{-1/2} e^{-\frac{1}{2}(x_i-u_j)^T(W+\Upsilon)^{-1}(x_i-u_j)}$$

for the value u_j for which we want the prediction of Y_{u_j} given the observed data, $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. We also incorporate outcome measurement noise as in chapter three, by using covariance $K = \Sigma + \sigma_\xi I$, where Σ is derived from C_n similarly as in equation 3.6. It has been shown that accounting for the additional noise from measurement process significantly enhances the learning process *and* subsequent prediction, but the estimation of the covariance form of eq. 6.12 is more challenging due to additional parameters (Dallaire, Besse & Chaib-draa, 2009), that is, due to need to estimate noise covariances Υ. As both, W and Υ, contain parameters to estimate, their products and sums induce trade-offs to likelihood function, resulting in multiple local maximas. This problem can be alleviated either by making strong prior assumptions for the model parameters (as in Dallaire, Besse & Chaib-draa, 2009) or by more elaborate approach presented in the previous chapters of current work. Here, all of the parameters have already been estimated/learned via developments of previous chapters, and we use Girard's approximative approach only to smoothing and prediction.

6.2 Interpreting the Squared exponential

As an afterthought for the current chapter, we now digress a bit to theoretical material of chapter 3. It turns out that we can give an interpretation to the Squared exponential covariance function, using Product of Gaussians lemma. Recall that there is an orthogonal expansion for the covariance function (Mercer's theorem). For the Squared exponential, it can be found in practice by starting from *radial basis functions* and using heuristics (MacKay, 2003). Consider taking Gaussian i.i.d. prior $\xi \sim N_k(0, \lambda I)$ over finite set, $\mathfrak{H}$, of indices on the real-interval $[h_{min}, h_{max}]$, and modeling $y = f(x)$ with a nonlinear regression using k radial basis functions,

$$f(x) = \sum_{h \in \mathfrak{H}} \xi_h \psi_h(x), \qquad \psi_h(x) := e^{-\frac{(x-h)^2}{2r^2}},$$

where $\{h\}_{h \in \mathfrak{H}}$ are some constants in the interval. Then, the covariance function is of the form

$$C(x, x') = E[f(x)f(x')] = \sum_h \sum_{h'} E[\xi_h \xi_{h'}] \psi_h(x) \psi_{h'}(x') = \lambda \sum_h \psi_h(x) \psi_h(x').$$

If we now fill the interval $[h_{min}, h_{max}]$ densely, letting $\lambda = S/\Delta h$ scale as the number of basis functions per unit interval, Δh, the sum can be thought of as

an integral

$$C(x, x') = S \int_{h_{min}}^{h_{max}} e^{-\frac{(x-h)^2}{2r^2}} e^{-\frac{(x'-h)^2}{2r^2}} dh.$$

By now taking $h_{max} \to \infty$ and $h_{min} \to -\infty$, and applying similar trick as for derivation of noisy covariance above, we gain some insight. That is, by using Product of Gaussians lemma and integration with respect to h, we find that

$$C(x, x') = \sqrt{\pi r^2} S e^{-\frac{(x-x')^2}{4r^2}}.$$

Using re-parameterization to more familiar terms, we find that this is nothing but the Squared exponential covariance in one dimensional case, that is,

$$C(x, x') = v e^{-\frac{1}{2}\frac{(x-x')^2}{W}}.$$

In general, it is not a trivial feat to be able to perform regression with infinite Gaussian shaped basis functions without overfitting to observations. With current constructions, it seems that this can be done in spite of the noise in predictor (independent) variables.

7 Summary and afterword

We have now derived a flexible non-parametric regression framework for outcome prediction using noisy behavioral measurements. "Noisy" meaning that desired values cannot be observed without significant error. The suggested plan of action is the following. Collect data and simultaneously estimate Factor analysis and GPR model, using Markov chain Monte Carlo Stochastic gradient estimation method. Here, we need the stochastic approach because we do not directly observe the *latent variable* of the Factor analysis model. This way, in spite of not observing index variable of GPR, we are able to estimate GPR model.

Once we have estimated GPR model, we still cannot use it for prediction via conditioning, because we do not have observations from the correct index variable. To circumvent this short-coming, use results from Factor analysis to derive a measurement scale, and its error properties. Then, set-up (chapter six) an approximating Gaussian process with covariance function that is otherwise identical, but *models* the index-variable noise as well as the outcome noise. Here, the measurement scale response is used as the indexing variable, for which we know the error properties. With this setting, we are able to perform statistical inference in the infinite-dimensional function space, whose elements are functions from *unobserved* latent variable to real-valued outcome variable.

While this is a natural point to end this mathematics thesis, it is obvious that above developments open up a much larger research program. For one thing, practical aspects of the method require a lot more attention. In spite of the theoretical convergence proofs, it is not obvious how effective stochastic gradient approach is in this context. In addition, we made somewhat susceptible approximation with another Gaussian process, having the so called "noisy covariance function". While simulation experiments indicate that this approximation works well (Girard, 2004; Dallaire, Besse & Chaib-draa, 2009), topic hardly can be considered as thoroughly researched. Furthermore, it is fairly safe to say that few (if none) have previously considered what would be optimal covariance structure for outcome process indexed by some behavioral construct (psychological, psychiatric, sociological, or other such, theory). To give an example, it is relatively open question how atherosclerosis risk varies as a function of human temperament traits (Hintsanen, et al., 2009). The fact that one needs to build everything around the measurement noise in indices considerable limits the available practical options for the covariance form, provided that one requires for empirical verification (as one should). Still, room

for research on covariance form exist, starting from linear covariance, which is also analytically tractable similarly to the Squared exponential (Girard, 2004).

If there would be growing interest to GPR approach within the behavioral research fields, this might open interesting possibilities in the field of *experimental design*, as discussed in the end of chapter three. If one could derive knowledge about the appropriate covariance function, and a link to a Markov-associated process, it might be possible to gain information on the optimal design prior the data collection (in the sense of minimizing uncertainty). This might help in the frequently encountered situation when collecting outcome observations is "expensive", economically or humanly. This kind of research must be seen, at best, as a fairly distant future. Reliable information on covariance forms would be required. However, there are other directions in the immediate near future that should be taken, for GPR approach to ever get this far. Namely, these are the very same steps that have led to popularity of GPR in the Machine learning and Geostatistics communities. One should show that it is effective in practice, beyond other, conceptually simpler, approaches.

What direction should be taken next in the research on Gaussian process approach to behavioral prediction? In my opinion, the best way to collect interest on the topic, and thereby needed resources, would be to show that it convincingly outperforms the standard linear (and perhaps quadratic) regression approaches, using fairly large data set and cross-validation (Arlot & Celisse, 2010). Reason to favor validation data -based methods over other model selection methods, is that they provide results seemingly independently of model assumptions. As we are speaking of prediction, it is important to ensure that we truly are able to predict new data points that were not used in model estimation. We conclude the current treatment of topic to these considerations.

Bibliography

Adler, R.J. and Taylor, J.E. (2007). Random Fields and Geometry. Springer-Verlag.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40-79.

Cai, L. (2010). High-dimensional exploratory Item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, 75, 1:33-57.

Cappé, O., Moulines, E. and Rydén, T. (2005) Inference in Hidden Markov Models. Springer, New York, USA.

Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C.M. (2006) Measurement Error in Nonlinear Models: A Modern Perspective (2nd Ed.). Chapman & Hall/CRC, Boca Raton, USA.

Cloninger, C.R., Przybeck, T.R., Svrakic, D.M. and Wetzel, R.D. (1993). The Temperament and Character Inventory (TCI): a guide to its development and use. Center for Psychobiology of Personality, Washington University ST. Louis (Mo).

Costa, Jr. P.T. and McCrae, R.R. (1985). The NEO-PI personality inventory manual. Odessa, FL: Psychological Assessment Resources.

Cudeck, R. and MacCallum, R.C. (Eds.) (2007), Factor Analysis at 100: Historical Developments and Future Directions. Lawrence Erlbaum Associates, Inc., Mahwah, New Yersey, USA.

Dallaire, P., Besse, C. and Chaib-draa, B. (2009). Neural Information Processing: Lecture Notes in Computer Science, 2009, C.S. Leung, M. Lee J.H. Chan (Eds.), Volume 5863/2009, 433-440, DOI: 10.1007/978-3-642-10677-4_49

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, 1:1-38.

Ellis, S.P. (2004), Instability of Statistical Factor Analysis, *Proceedings of the American Mathematical Society*, 132, 6:1805-1822.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions in Pattern analysis and Machine Intelligence*, 6, 721-741.

Girard, A. (2004). Approximate Methods for Propagation of Uncertainty with Gaussian Process Models. A thesis submitted to the University of Glasgow for the degree of Doctor of Philosophy. Available in www.dcs.gla.ac.uk/ rod/publications/Gir04.pdf

Gu, M.G. & Zhu, H. (2001). Maximum likelihood estimation for spatial models by Markov chain Monte Carlo Stochastic approximation. *Journal of the Royal Statistical Society-Series B*, 63, 339-355.

Hewitt, E. & Ross, K.A. (1997). Abstract Harmonic Analysis, Volume II: Structure and Analysis for Compact Groups, Analysis on Locally Compact Abelian Groups. Springer-Verlag, Berlin, Germany.

Hintsanen, M., Pulkki-Råback, L., Juonala, M., Viikari, J.S.A., Raitakari, O.T., Keltikangas-Järvinen, L. (2009) Cloninger's temperament traits and early atherosclerosis: The Cardiovascular risk in young Finns study. *Journal of Psychosomatic Research*, 67, 1:77-84.

Jennrich, R.I. (2007). Rotation methods, algorithms, and standard errors. Factor Analysis at 100, Historical Developments and Future Directions (Chapter 15, Eds. Cudeck, R. & MacCallum, R.C.), Lawrence Erlbaum Associates, Inc., Mahwah, New Yersey, USA.

John, O.P., Robins, R.W. & Pervin, L.A. (Eds.) (2008). Handbook of Personality: Theory and Research (third edition). The Guilford Press, New York, USA.

Klenke, A. (2008) Probability Theory: A Comprehensive Course. Springer-Verlag, London, UK.

Lawley, D.N. and Maxwell, M.A. (1971), Factor Analysis as a Statistical Method. Butterworth & Co, London, England.

Lopes, H.F. and West, M. (2004) Bayesian model assessment in Factor analysis. *Statistica Sinica*, 14, 41-67.

MacKay, D.J.C. (2003). Information theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge, UK.

Magnus, J.R. and Neudecker, H. (1991). Matrix Differential Calculus with Applications in Statistics and Econometrics, Revised Edition. John Wiley & Sons, Chippenham, UK.

Marcus, M.B. and Rosen, J. (2006). Markov processes, Gaussian processes, and Local times. Cambridge University Press, New York, USA.

Martin, J.K. and McDonald, R.P. (1975). Bayesian estimation in unrestricted Factor analysis: A treatment for Heywood cases. *Psychometrika*, 40, 4:505-517.

Moustaki, I (2007). Factor analysis and latent structure of categorical and metric data. Factor Analysis at 100: Historical Developments and Future Directions (Chapter 14, Eds. Cudeck, R. & MacCallum, R.C.), Lawrence Erlbaum Associates, Inc., Mahwah, New Yersey, USA.

Plomin, R., Haworth, C.M.A and Davis, O.S.P. (2009). Common disorders are quantitative traits, *Nature Reviews Genetics*, 10, 872-878.

Puttonen, S., Elovainio, M., Kivimäki, M., Koskinen, T., Pulkki-Råback, L., Viikari, J.S.A., Raitakari, O.T. and Keltikangas-Järvinen, L. (2008). Temperament, health-related behaviors, and automatic cardiac regulation: The cardiovascular risk in young Finns study. *Biological Psychology*, 78, 2:204-210.

Rasmussen, C.E. (1996). Evaluation of Gaussian processes and other methods for non-linear regression, PhD thesis, University of Toronto.

Rasmussen, C.E. and Williams C.K.I. (2006). Gaussian Processes for Machine Learning. The MIT Press, Massachusetts, USA.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, 22, 400-407.

Rue, H. and Held, L. (2005) Gaussian Markov Random Fields: Theory and Applications. Chapman & Hall, Boca Raton, USA.

Tarkkonen, L. and Vehkalahti, K. (2005). Measurement errors in multivariate measurement scales. *Journal of Multivariate Analysis*, 96, 172-189.

Wu, C.F.J. (1983). On the convergence properties of the EM-algorithm. *The Annals of Statistics*, 11, 1:95-103.

Ylvisaker, D. (1987). Special invited paper: Prediction and design. *The Annals of Statistics*, 15, 1:1-19.