Probabilistic Models for Human Behavior Learning

Autor: Pablo Moreno-Muñoz

Tesis depositada en cumplimiento parcial de los requisitos para el
Grado de Doctor en Multimedia y Comunicaciones

Universidad Carlos III de Madrid

Director y tutor: Antonio Artés Rodríguez

March 2021

ACKNOWLEDGEMENTS

En estos años dedicados al doctorado, puedo decir con orgullo que he aprendido a calcular, resolver integrales, formular, aproximar y estimar casi cualquier cosa que se represente por una letra griega, pero nada de ello me ayuda a la hora de resolver la ecuación más difícil de todas. Esta es, de alguna forma, el cómo agradecer el apoyo de todas las personas que me han acompañado durante este tiempo. Adelanto que pocas no son. Y no negaré que nunca haya pensado en cómo hacer esto con anterioridad, pero supongo que simplemente ha llegado el momento. Quienes me conocéis bien sabéis que no soy mucho de formalidades estériles. Allá voy.

Digamos que sí, de alguna manera la investigación llamó a mi puerta y cómo siempre me han dicho; a veces, los caminos le eligen a uno. Confieso aquí que los cantos de sirena de un doctorado internacional me llamaron, con fuerza, pero fue mi director Antonio Artés quién en ese proceso de confusión que cualquier estudiante recién salido del horno (o máster) siente, me dió una oportunidad de hacer aquello que me gustaba en mi hogar, cerca de mi familia y mi vida. La moneda de cambio fue el poder realizar estancias, y tener esa experiencia que para mí era fundamental. Puedo asegurar que cumplió su promesa. Por ello, quiero agradecerle todo su apoyo durante estos años, todo lo que he aprendido y las oportunidades que me ha brindado. Sin su confianza en mi trabajo nada de esto sería posible ni podría haberme formado cómo lo he hecho. La deuda en este sentido es simplemente, incalculable. Continúo y debo agradecer a David Ramírez, de quién he aprendido el valor del rigor en esta profesión que es ser investigador, así cómo le agradezco su paciencia y gran labor por *masticar* aquello que fue mi primer paper, hoy afortunadamente aceptado en una revista. Valoro su apoyo durante estos años, y esta tesis no sería la misma sin su contribución. Gracias también a Mauricio Álvarez por su infinita ayuda en mis fabulosos meses en Sheffield, si puedo decir hoy algo sobre Gaussian processes, es gracias a él. Quién diría que nuestra primera conversación en el laboratorio 4.3.B01 durante su visita en 2017 conllevó el punto de inflexión de mi doctorado, y una estrecha colaboración que a día de hoy continúa.

Entrando aún más en materia, quiero de alguna manera dejar aquí constancia del gran viaje que ha sido este doctorado y dar por tanto las gracias a todos aquellos que me habéis acompañado. Comienzo por los *mayores*, aquellos que cuando entré en 2016 cómo estudiante de máster en el laboratorio me aconsejaron y ayudaron en todo durante mis primeros pasos. Ellos son Melanie Fernández, Deniz Akyildiz, Gonzalo Ríos, Alfredo Nazábal, Grace Villacrés y Fran Hernando, con quienes compartí muy buenos momentos tanto fuera cómo dentro de la universidad. Para mí fueron la generación que habría camino por delante de mí, y hoy me enorgullece ver sus logros. Quiero también mencionar el recuerdo de los magníficos seminarios semanales del grupo de investigación, punto de encuentro con Gonzalo Vázquez, Toby Koch, Pablo M. Olmos, Joaquín Míguez, Javier López y Luca Martino, todos grandes profesores e investigadores por los que mi admiración siempre ha sido enorme. Me llevo de todos ellos grandes enseñanzas, tanto técnicas, cómo de la vida investigadora y muy buenos recuerdos de largas comidas y cenas de grupo. No me olvido de Ana Hernando, quién en los momentos complicados siempre me ayudó y fue el apoyo que evitó mi desesperación.

Reconozco que han sido muchos congresos, summer schools, viajes y vuelos (probablemente serán más en el futuro si la *maldita* pandemia deja de acompañarnos), en los que he conocido gente maravillosa de esta gigantesca comunidad que es hoy el *machine learning*, pero

quiero hacer hincapié sobretodo en mi agradecimiento a vosotros, quienes me acompañasteis en mis aventuras en UK y Alemania. Following the temporal order, I should begin with the amazing internship in Sheffield during 2018. Thanks to Wil Ward (the second "ℓ" was lost in a fight), for his continuous support during that fabulous months, for his friendship, for his cooking habilities and for being the best mate during our glorious trip to Canada for the NeurIPS Conference in Canada in December 2018. Also thanks to Fariba Yousefi, Alasdair Warwicker, George Hall, my desk-mate Senee Kitimoon and the rest of member of the lab for their support. Gracias también a Juan José Giraldo por su apoyo, compañía y largos ratos de café, si hoy mi cariño por Colombia es inmenso, gran parte es por él. Sigo por aquella primavera en Tübingen en 2019, por lo que significó la experiencia y los que estuvieron a mi lado. Gracias a Isabel Valera por hacer aquello posible, por todo lo que aprendí, y si hoy la perspectiva *humana* en esta carrera profesional es prioritaria para mí es porque ella me lo inculcó. Agradezco también a Adrián Javaloy su fiel compañía, su energía y habilidad a lomos de las matemáticas y el *software*, y dejo aquí un feliz recuerdo de los días en los que el proyecto Hölder se expandía sin freno por las pizarras de tiza del instituto. I am sure that he, Martina Contisciani, Nicolò Ruggeri and Andrea della Vecchia have good memories of our long days and nights in the old city of Tübingen, thanks for your friendship at that times. Also thanks to Amir Hossein Karimi, for his honest happiness, and the rest of people from the Empirical Inference Department. No me olvido de Diego Agudelo, con quién compartí mucho más que mesa en aquellos meses del Max Planck. Sin su apoyo y ayuda constante, mi tiempo allí no hubiera sido el mismo de ninguna forma, por lo que estaré siempre agradecido.

Por evitar un exceso edulcorante en estas líneas, quiero también *no* agradecer a la nevada maligna (prima-hermana de Filomena) aquella llegada caótica a Sheffield en Enero 2018, con cursillo avanzado de transbordo de trenes sobre el hielo; y el coraje de aquellas personas que dejaron que estuvieramos 21 días sin calefacción ni agua caliente en un Febrero alemán por culpa de un concurso público para contratar 5k litros de gasoil. Estoy seguro de que Kamil Adamczewski guarda un *no* feliz recuerdo de aquellas noches gélidas. A él le agradezco no quemar la casa con el secador, su fuente favorita de calor, ser el políglota más increíble que he conocido y sobretodo su alegría contagiosa.

Avanzo ya a los instantes finales de este *breve* reconocimiento en forma de recuerdos, y menciono a aquellos con los que más he compartido momentos. Comienzo por Ignacio Peis, el granaíno de Jaén con quién he recorrido media Europa en infinidad de aventuras, de paciencia inquebrantable, compañero de largas filosofías, historias inconfesables y sobretodo un gran amigo. Prosigo con Juan José Campaña, a quién agradezco su apoyo fundamental en estos años, admiro su valía, y con quién he crecido codo con codo tanto personal cómo profesionalmente. Gracias también a Lorena Romero, Fernando Moreno y Daniel Barrejón por los buenos momentos juntos, por construir vida fuera de la universidad y hacerme sentir útil con mis consejos por los que otros no darían un duro. Dejo aquí un recuerdo de agradecimiento también a Pablo Sanchez, valiente cómo pocos, de quién admiro su inteligencia y cuya marcha a Alemania fue un momento agridulce para Peis y para mí. Gracias también a Emese Sukei y Patricia Carretero por ser siempre un ejemplo de trabajo infatigable y amabilidad constante. No me olvido de aquellos que me han aportado mucho más de lo que piensan; mis antiguos alumnos del máster, Ángela Moreno, Jesús Herrera, María Martínez, Alex Guerrero y Óscar Jímenez, hoy ya la nueva generación de investigadores en el laboratorio, en cuyos ojos vi la ilusión pura por aquello a lo que he dedicado tanto tiempo y esfuerzo. Eso me dió fuerzas para seguir adelante y creer aún más en lo que hago. El futuro es vuestro.

Termino por las personas fundamentales en mi vida, más allá de lo que estos años de doctorado han sido. Gracias a Carlos García por su longeva amistad, hacerme ser mejor persona, tratar mis adicciones (azúcar, café, ...), ser experto diseñador de motes y sobretodo siempre estar ahí. También a Jaime Castilla, brillante arquitecto y artista, cuya compañía se

cuenta ya en décadas, y cuya bondad y amistad siempre estuvo cuando la necesité. Añado a Juan Prieto, de quién admiro su inteligencia y razón, agradezco su siempre atento apoyo y cuyo suplicio opositor (hoy logrado) me acompañó en los largos años de tesis. Gracias también a Jimena López, por conservar nuestra amistad adolescente, ser la auténtica felicidad libre y dar voz a mi conciencia. No puedo olvidarme de Javier Martínez, co-autor de la foto que encabeza este documento y que a todos los efectos es el hermano que nunca tuve, si no me encontráis, seguramente esté recorriendo cordilleras con él. Añado también a Pablo Díez e Isa López, mis grandes amigos de la carrera y con los que pasé algunos de los momentos más felices de mi vida; a ellos les agradezco su amistad y haber sido partícipes de aquello que construimos.

Por último, gracias a mis padres, Marisa y Luis, cuyos apellidos llevo unidos con orgullo en reconocimiento a su esfuerzo inagotable por mi crecimiento y mi felicidad. Hoy acabo más de 20 años de formación (que no aprendizaje), desde que entré en el Colegio Estudio por primera vez, y el mérito es tanto mío como suyo. Mi pasión por la ciencia, el conocimiento y la investigación viene de ellos y de sus logros que admiro; pocos entenderán mejor lo que significa esta tesis para mí.

Y fin, fin contigo, quién diste y das luz a mi vida, hoy representas todo lo que uno puede esperar de la felicidad y me acompañas en este final. Juntos, el futuro es nuestro.

Os espero a todos en Copenhague.

CONTENIDOS PUBLICADOS Y PRESENTADOS

La siguiente relación de trabajos es una bibliografía corta de artículos de revista, conferencia, workshops y *preprints*, incluidos como parte de esta tesis. En cada contribución está indicado su presencia *completa* o *parcial* en esta tesis. Además, su presencia en los capítulos se indica en cada uno de los párrafos introductorios de los mismos. Por último declaro: *El material de esta fuente incluido en la tesis no está señalado por medios tipográficos ni referencias.*

Conferences & Workshops.

1. P. Moreno-Muñoz, A. Artés and M. Álvarez. Heterogeneous multi-output Gaussian processes. *In Advances in Neural Information Processing Systems (NeurIPS)*, 2018. (spotlight) [pdf] – Presencia *completa* en Ch. 2.
2. P. Moreno-Muñoz, D. Ramírez and A. Artés. Continual learning for infinite hierarchical change-point detection. *In International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020. [pdf] – Presencia *parcial* en Ch. 4.
3. L. Romero-Medrano, P. Moreno-Muñoz and A. Artés. Multinomial sampling for hierarchical change-point detection. *In International Workshop on Machine Learning for Signal Processing (MLSP)*, 2020. [pdf] – Presencia *parcial* en Ch. 4.
4. P. Moreno-Muñoz, A. Artés and M. Álvarez. Recyclable Gaussian processes. *In arXiv preprint arXiv:2010.02554 (to be re-submitted)*, 2020. [pdf] – Presencia *completa* en Ch. 3.
5. P. Moreno-Muñoz, L. Romero-Medrano, A. Moreno, J. Herrera-López, E. Baca-García and A. Artés. Passive detection of behavioral shifts for suicide attempt prevention, *Machine Learning for Mobile Health Workshop (ML4MH) at NeurIPS*, 2020. [pdf] – Presencia *parcial* en Ch. 5.

Journals.

6. S. Berrouiguet, D. Ramírez, M. L. Barrigón, P. Moreno-Muñoz, R. Carmona, E. Baca-García and A. Artés-Rodríguez. Combining continuous smartphone native sensors data capture and unsupervised data mining techniques for behavioral changes detection: A case series of the Evidence Based Behavior (eB2) study, *In Journal of Medical Internet Research (JMIR)*, 2018. [pdf] – Presencia *parcial* en Ch. 5.
7. P. Moreno-Muñoz, D. Ramírez and A. Artés. Change-point detection in hierarchical circadian models. *In arXiv preprint arXiv:1809.04197 (accepted in Pattern Recognition)*, 2018. [pdf] – Presencia *parcial* en Ch. 2 and Ch. 4.
8. P. Moreno-Muñoz, A. Artés and M. Álvarez. Continual multi-task Gaussian processes. *In arXiv preprint arXiv:1911.00002 (second review round in JMLR)*, 2019. [pdf] – Presencia *parcial* en Ch. 3.

ABSTRACT

THE problem of human behavior learning is a popular interdisciplinary research topic that has been explored from multiple perspectives, with a principal branch of study in the context of computer vision systems and activity recognition. However, the statistical methods used in these frameworks typically assume short time scales, usually of minutes or even seconds. The emergence of mobile electronic devices, such as smartphones and wearables, has changed this paradigm as long as we are now able to massively collect digital records from users. This collection of smartphone-generated data, whose attributes are obtained in an unobtrusive manner from the devices via multiple sensors and *apps*, shape the behavioral footprint that is unique for everyone of us. At an individual level, the data projection also differs from person to person, as not all sensors are equal, neither the *apps* installed, or the devices used in the real life. This point actually reflects that learning the human behavior from the digital signature of users is an arduous task, that requires to fuse *irregular* data. For instance, collections of samples that are corrupted, heterogeneous, outliers or have short-term correlations. The statistical modelling of this sort of objects is one of the principal contributions of this thesis, that we study from the perspective of Gaussian processes (GP).

In the particular case of humans, as well as many other life species in our world, we are inherently conditioned to the *diurnal* and *nocturnal* cycles that everyday shape our behavior, and hence, our data. We can study these cycles in our behavioral representation to see that there exists a perpetual *circadian* rhytm in everyone of us. This *tempo* is the 24h periodic component that shapes the baseline temporal structure of our behavior, not the particular patterns that change for every person. Looking to the trajectories and variabilities that our behavior may take in the data, we can appreciate that there is not a single repetitive behavior. Instead, there are typically several patterns or *routines*, sampled from our own dictionary, that we choose for every special situation. At the same time, these routines are arbitrary combinations of differents timescales, correlations, levels of mobility, social interaction, sleep quality or will for working during the same hours on weekdays. Together, the properties of human behavior already indicate to us how we shall proceed to model its structure, not as unique functions, but as a dictionary of latent behavioral profiles. To discover them, we have considered latent variable models.

The main application of the statistical methods developed for human behavior learning appears as we look to medicine. Having a personalized model that is accurately fitted to the behavioral patterns of some patient of interest, sudden changes in them could be early indicators of future relapses. From a technical point of view, the traditional question use to be if newer observations conform or not to the expected behavior indicated by the already fitted model. The problem can be analyzed from two perspectives that are interrelated, one more oriented to the characterization of that single object as outlier, typically named as *anomaly detection*, and another focused in refreshing the learning model if no longer fits to the new sequential data. This last problem, widely known as *change-point detection* (CPD) is another pillar of this thesis. These methods are oriented to mental health applications, and particularly to the *passive* detection of crisis events. The final goal is to provide an early detection methodology based on probabilistic modeling for early intervention, e.g. prevent suicide attempts, on psychiatric outpatients with severe affective disorders of higher prevalence, such as depression or bipolar diseases.

RESUMEN

El problema de aprendizaje del comportamiento humano es un tema de investigación interdisciplinar que ha sido explorado desde múltiples perspectivas, con una línea de estudio principal en torno a los sistemas de visión por ordenador y el reconocimiento de actividades. Sin embargo, los métodos estadísticos usados en estos casos suelen asumir escalas de tiempo cortas, generalmente de minutos o incluso segundos. La aparición de tecnologías móviles, tales cómo teléfonos o relojes inteligentes, ha cambiado este paradigma, dado que ahora es posible recolectar ingentes colecciones de datos a partir de los usuarios. Este conjunto de datos generados a partir de nuestro teléfono, cuyos atributos se obtienen de manera no invasiva desde múltiples sensores y *apps*, conforman la huella de comportamiento que es única para cada uno de nosotros. A nivel individual, la proyección sobre los datos difiere de persona a persona, dado que no todos los sensores son iguales, ni las *apps* instaladas así cómo los dispositivos utilizados en la vida real. Esto precisamente refleja que el aprendizaje del comportamiento humano a partir de la huella digital de los usuarios es una ardua tarea, que requiere principalmente fusionar datos *irregulares*. Por ejemplo, colecciones de muestras corruptas, heterogéneas, con *outliers* o poseedoras de correlaciones cortas. El modelado estadístico de este tipo de objetos es una de las contribuciones principales de esta tesis, que estudiamos desde la perspectiva de los procesos Gaussianos (GP).

En el caso particular de los humanos, así como para muchas otras especies en nuestro planeta, estamos inherentemente condicionados a los ciclos diurnos y nocturnos que cada día dan forma a nuestro comportamiento, y por tanto, a nuestros datos. Podemos estudiar estos ciclos en la representación del comportamiento que obtenemos y ver que realmente existe un ritmo *circadiano* perpetuo en cada uno de nosotros. Este *tempo* es en realidad la componente periódica de 24 horas que construye la base sobre la que se asienta nuestro comportamiento, no únicamente los patrones que cambian para cada persona. Mirando a las trayectorias y variabilidades que nuestro comportamiento puede plasmar en los datos, podemos apreciar que no existe un comportamiento único y repetitivo. En su lugar, hay varios patrones o *rutinas*, obtenidas de nuestro propio *diccionario*, que elegimos para cada situación especial. Al mismo tiempo, estas rutinas son combinaciones arbitrarias de diferentes escalas de tiempo, correlaciones, niveles de mobilidad, interacción social, calidad del sueño o iniciativa para trabajar durante las mismas horas cada día laborable. Juntas, estas propiedades del comportamiento humano nos indican cómo debemos proceder a modelar su estructura, no cómo functiones únicas, sino cómo un diccionario de perfiles ocultos de comportamiento, Para descubrirlos, hemos considerado modelos de variables latentes.

La aplicación principal de los modelos estadísticos desarrollados para el aprendizaje de comportamiento humano aparece en cuánto miramos a la medicina. Teniendo un modelo personalizado que está ajustado de una manera precisa a los patrones de comportamiento de un paciente, los cambios espontáneos en ellos pueden ser indicadores de futuras recaídas. Desde un punto de vista técnico, la pregunta clásica suele ser si nuevas observaciones encajan o no con lo indicado por el modelo. Este problema se puede enfocar desde dos perspectivas que están interrelacionadas, una más orientada a la caracterización de aquellos objetos como *outliers*, que usualmente se conoce como detección de anomalías, y otro enfocado en refrescar el modelo de aprendizaje si este deja de ajustarse debidamente a los nuevos datos secuenciales. Este último problema, ampliamente conocido como *detección de puntos de cambio* (CPD) es

otro de los pilares de esta tesis. Estos métodos se han orientado a aplicaciones de salud mental, y particularmente, a la detección *pasiva* de eventos críticos. El objetivo final es proveer de una metodología de detección temprana basada en el modelado probabilístico para intervenciones rápidas. Por ejemplo, de cara a preveer intentos de suicidio en pacientes fuera de hospitales con trastornos afectivos severos de gran prevalencia, como depresión o síndrome bipolar.

CONTENTS

CHAPTER 1

INTRODUCTION

HUMANS, as well as many other life species in our world, are inherently conditioned to the *diurnal* and *nocturnal* cycles that everyday shape our behavior. In the particular case of human individuals, such cycles can be accurately measured with modern technology, e.g. electronic devices, and they appear to be periodic every 24 hours. We can easily study these cycles to see that there exists a perpetual *circadian* rhythm in everyone of us. However, the existence of periodic components only indicates our *tempo* and hence, the baseline temporal structure that underlies our lifes, not the particular patterns that are different for every individual. With this idea in our minds, we can state that every single person has its own behavioral footprint, circadian itself, that only shares a few common pillars with the rest of the population, but remains still unique for her or him.

Looking to the particular paths and variabilities that our behavior may take, we appreciate that there is not any single repetitive behavior in everyone of us every day. Instead, there are typically several patterns or *routines*, that we choose for every special situation. At the same time, our routines are arbitrary combinations of different timescales, levels of *mobility*, social interaction, sleep or will for working during the same hours every weekday. Together, these properties of the human behavior already indicate to us how we shall begin to model its structure, not as an unique latent function but as a dictionary of latent behavioral profiles. We may discover them, if we look to the data that we all generate every single day.

This thesis advances in the problem of modelling the human behavior as a sequence of latent behavioral elements that we aim to discover based on uncertainty quantification. The main application of this sort of models appears as we look to medicine. Having a personalized model that is accurately fitted to the behavioral routines of some patient of interest, sudden changes in them can be early indicators of future relapses. This is of particular relevance in the context of mental health, where the assessment and passive monitoring of sufferers is a key milestone, for instance, to detect critical events for the evolution of chronic disorders. To achive this goal using machine learning (ML) methods, with the purpose of being later integrated in digital systems, we explore three main areas of study. First, we consider general models for the problem of learning from heterogeneous observations, that is, high-dimensional combinations of several statistical data types (i.e. real-valued, binary or categorical variables). We then study novel inference methods, in particular these ones that are capable of being deployed in *online*, or distributed scenarios. We pay a special attention to the problem of continually learn their parameters for being adapted to the evolution of behavior along time. Finally, our discoveries are connected to change-point algorithms, with the idea of designing general detectors for any type of human-generated data that we may be interested in.

1.1 Human Behavior Learning

The problem of modelling human behavior has been already explored from multiple perspectives, with a principal branch of study in the context of computer vision systems (Chaaraoui et al., 2012). The ideas behind this family of methods are typically focused in the identification of the short-term activity performed, i.e. sitting, walking, running, paying, etc, from raw

video sequences (Ma et al., 2016; Nunez et al., 2018) or simply images (Jalal et al., 2015). Others use information from inertial sensors (Nazabal et al., 2015) for automatic activity recognition, which is indeed a popular interdisciplinary research topic. It is worthy to mention some of their applications, among the ones, security, wellbeing systems and monitoring of elderly patients are significant to us. However, such learning methods often focus on a smaller timescale, usually seconds and/or minutes. Here, we are interested in the human behavior with important attention to the circadian component of the observations, which is in the order of hours and days.

The scope of the work presented in this thesis goes further than the previously mentioned methods, as we deal with the idea of personalized behavioral footprints that are projected on the smartphone-generated data. Also, on the records collected everyday in our electronic devices from multiple apps and sensors. At an individual level, this projection differs from person to person, as not all sensors are the same and neither the *apps* installed, or the devices used, which are never imposed. This purpose actually reflects that modelling human behavior from the digital signature of an individual is an arduous task, that requires to fuse *irregular* data. This idea of observations with *irregularities* will be later discussed in depth. The statistical modelling of this sort of objects is one of the principal contributions of this thesis.

The study of the digital circadian cycles of individuals is recursively visited in this work, where we based our initial perspective on the preliminary approach of Aledavood et al. (2015a). Their findings illustrated, in the beginning, that the circadian rhythms of people persist in the aggregated data from multiple sources, e.g. from mobility longitudinal traces, social interactions or even call registers. This motivated us to apply a probabilistic modelling on the top of the data aggregation, in order to capture the underlying components that might be hidden in every case. To do that, we got inspired by the idea of *eigenbehaviors* presented in Eagle and Pentland (2009), where an individual's behavior could be approximated as a weighted sum of latent components or vectors. Particularly, we found a coincidence in the timescale that we where looking for, in the order of hours to capture a latent structure valid for every day. Moreover, the work of Eagle and Pentland (2009) shared with the previous Eagle and Pentland (2006), the idea of multi-modal data modelled with discrete latent objects that we also adopted. However, this last approach focused in the analysis of complex social systems and particularly on the interactions between individuals, something that falls out of the scope of our project.

1.1.1 A Mental Health Perspective

The emergence of mobile electronic devices such as personal smartphones or wearables has also gained significant attention in healthcare due to their ubiquitous conditions, mainly for pervasive sensing. It is now widely known as *electronic* health (e-Health) in the literature. In particular, this disruption of mobile technologies afforded new opportunities (Miller, 2012) to obtain objective, reliable and real-time monitoring data of patients outside the ambulatory domains where assessment cannot be driven in a formal and daily manner.

For mental health, the principal advantage of such personal mobile devices is that their embedded monitoring systems are completely unobtrusive to the users, typically outpatients with chronic disorders. This property avoids direct interactions of clinicians with patients, that are often time-consuming and limit the potential counfounders due to the self-representation. Moreover, we are aware of the difficulties carried out by psychiatry clinicians to daily assess the state of patients. In this context, the degree of disability has been traditionally assessed using periodic reports written down by patients or their caregivers. However, this sort of protocols limit the utility of evaluations, as their are typically of poor

reliability or patients are unaware of their own symptoms that may lead to imminent relapses.

The presence of smartphones in patient's pockets, opens the door to novel methodologies, e.g. statistical behavior modelling for improving the assessment of the life conditions in chronic psychiatric outpatients (Osmani, 2015; Marzano et al., 2015; Firth et al., 2016; Barrigón et al., 2017). Mainly, those ones with diseases of higher prevalence, such as schizophrenia or affective mood disorders (depression or bipolar diseases). To help on these clinical tasks, we find a key point of connection between behavioral modelling and new methods on e-Health. Our final idea for the application of the statistical methods is therefore to provide a digital support to psychiatric clinicians, within useful tools for the systematic monitoring and assessment of the behavioral state of patients during their daily lifes out of hospitalary domains.

1.1.2 The Role of Change Detection

Once a probabilistic model is well fitted to the clinical data of interest, or in our particular case, to the behavioral observations monitored from patients, we do think about its potential usage. As machine learning practitioners, using statistical models (without supervision) for clinical diagnosis is a meaningless strategy. We cannot deliver to digital systems the decision-making on sensitive chronic patients. Instead, and also based on the data-driven discipline chosen for this thesis, we can still detect if new data conform or not to the expected behavior indicated by the model. In other words, we should not say if some behavioral pattern is *beneficial* or not for the the disease, but we can detect if something is just different that the events previously observed. If there is a change at some time step of the sequential data, we can detect it.

From a medical point of view, the appearance of changes has deserve some attention. This detection of changes, often considered as relapses for early intervention systems have demonstrated to be of practical use in mental health, particularly in schizophrenia (Barnett et al., 2018; Torous et al., 2017). Our aim in this thesis is to explore this theme, providing a full detection tool for clinicians with three key properties. First, to approximate every patient's behavioral routines in a reliable manner using statistical components with some degree of interpretability. Second, having statistical models that discriminate if new data belongs or not to the learned representation, we develop a change-point detection method where outputs are also interpretable in a medical context. Finally, based on the recent spirit of e-Health and also focused in the mental health problem, we identify relevant contributions of the aforementioned statistical methods for suicide attempt prevention.

1.1.3 Problems of Behavioral Data Modelling

We have partially shown our purpose for modelling human behavior via statistical methods based on uncertainty quantification. However, if we look to the applicability of such methods in real-world examples, most attention has been payed to the development of complex learning systems that only work on *ultra preprocessed* data. We do refer to this type of data as the case of having all samples in the same statistical domain, i.e. continuous or discrete variables, without the presence of outliers or having filtered out the missing attributes or objects that overflow the sensing range. As a consequence, two main problems typically affect this sort of preprocessing stages. The first issue is related to the growth of datasets, whose observations already come from thousands to millions. Such filtering processess use to be more and more time-demanding and, in some scenarios with temporal constraints, it is not possible to filter the data in an affordable time. The second issue emerges from a statistical point of view.

Filtering out the undesired data is not a realistic solution for models and might also lead to biased results or even non-understandable conclusions in many cases.

Fortunately, probabilistic ML methods have demonstrated to be more robust to non-preprocessed data than other deterministic methods, for instance, frequentist or discriminative models. If some of the features are lost, we can still infer them conditionally, given the rest of information, or if there is plenty of outliers, we may know how unlikely they are, having the observed data. Note that this paradigm is specially well suited to the applications considered in this thesis, as human behavior is typically projected into an *irregular* multimodal data that is correlated among dimensions.

Irregular Observations

We know that there is extensive work in the literature about probabilistic learning methods for *regular* data, and more specifically, for *homogeneous* observations. We refer to these type of data as the set of objects whose attributes are assumed to be always of the same statistical nature. This assumption often leads to the choice of a single likelihood density, e.g. Gaussian, Bernoulli, categorical, etc, that links the statistical uncertainty of the data within the latent structure of the problem via the parameters. If we look to the particular problem of human behavior learning, this assumption is not possible to be taken. As said above and also based on the empirical expirience, behavioral data from smarpthones is typically generated by different sensors, *apps* or contains very different information. In addition to the problem of being a temporal problem, where an outcome should be presented as soon as possible, then, transforming all the attributes into a common statistical codification is a handicap. However, our reference to *irregular* data does not finish with the heterogeneous components, as we also consider high-dimensional objects, with outliers and even periodic correlations between features at different scales.

Flexible Inference Methods

We have mentioned how statistical learning systems must face *irregular* datasets with heterogeneous attributes to estimate human behavior. However, a second drawback also limits the direct application of state-of-the-art methods in this context. Due to the human behavioral data shape an underlying changing representation, that often evolves along time, we must consider temporal statistical models for this challenge. This temporal evolution of the observations opens the door to two strategies. The first one would be to wait until a sufficient amount of data records were stored, and then, use the statistical learning method to model both the behavior and its evolution along time. The main problem of this point of view is that neither clinicians or patients can wait years for a predictive output. Instead, the second strategy addresses the temporal correlation in a different way. The idea is to adapt models in a recursive manner, that is, using *online* methods. This has motivated many researchers to consider *online* settings in the literature, but here we face the challenge within the *irregular* data conditions. For this reason, we adopt the notion of having *flexible* inference processes that could be adapted to sequential observations as well as distributed setups for preserving the privacy of patients.

1.2 Overview of Models and Contributions

This thesis is oriented to the design, implementation and evaluation of statistical learning models, based on probability theory and uncertainty quantification, for the problem of human behavior learning. Having reliable representations of an individual behavior is the first step

in order to detect abnormal events or shifts that can be harbingers of imminent relapses in the context of medicine. For this task, we have faced several problems during the development of this thesis. The proposed statistical methods for solving them are also the principal contributions of our work. We now survey the ones proposed for the modelling task, the inference challenges and the potential applications that we found in mental health.

1.2.1 Gaussian Processes for Heterogeneous Data

There is remarkable evidence that by simultaneously exploiting the correlations between multiple attributes in observed objects, it is possible to provide better predictions, particularly in scenarios with missing or noisy data. In the context of Gaussian process (GP) models, both Bonilla et al. (2008); Dai et al. (2017) have demonstrated this statement. With this growing interest, the apparition of multi-output GPs generalised the powerful predictive model to the vector-valued setup (Alvarez et al., 2012). Regarding the type of output data to be modelled, the main focus in the literature for GPs has been on regression problems for continuous variables. Traditionally, outputs are assumed to follow a Gaussian likelihood distribution where the mean function is parameterized by the GP and the variance is often treated as unknown hyperparameter. Bayesian inference is tractable for these models.

However, few attempts have tried to extend the GP setting, and particularly the MOGP, to other types of likelihoods. For examples, Skolidis and Sanguinetti (2011) used the output functions of a MOGP for jointly modelling several binary classification problems, each of which used a *probit* likelihood. In this case, posterior inference was performed using expectation-propagation (EP) and the variational mean-field approach. Both Chai (2012) and Dezfouli and Bonilla (2015) used the coregionalization model in MOGPs for modeling a *single categorical variable* with multinomial logistic likelihoods. The output variables in this model were used as replacements for the linear predictors in the softmax function. For the single-output GP case, the usual practice for handling non-Gaussian data has been replacing each parameter of the likelihood model by one or more *independent* GP priors. Since computing integrals becomes often intractable, different alternatives have been offered for approximate inference. Examples are the Gaussian heteroscedastic regression model (Lázaro-Gredilla and Titsias, 2011), Laplace approximations (Vanhatalo et al., 2013), Poisson likelihoods (Saul et al., 2016) or even Wishart processes (Wilson and Ghahramani, 2011).

Our contribution in this area of study is to provide an extenson of MOGP models for prediction in heterogeneous databases. This is the case for which the output variables are an arbitrary combination of continuous, categorical, binary or other discrete variables, each one with a different likelihood distribution. The key principle is to use the outputs of the MOGP as the latent functions that *modulate* the parameters of several likelihood functions, one likelihood density per heterogeneous attribute. Based on the inducing variable formalism in MOGPs introduced by Alvarez and Lawrence (2009), we introduce variational methods for inference. We experimentally provide evidence of the benefits of simultaneously modeling heterogeneous data in different applied problems, among the ones, human behavior modelling stands out. Interestingly, our implementation of the model follows the spirit of Hadfield et al. (2010), where the user of the learning system only needs to specify a list of desired likelihood densities: `likelihood_list = [Bernoulli(), Poisson (), HetGaussian()]`, where `HetGaussian()` refers to the heteroscedastic Gaussian distribution, and the number of latent parameter functions per likelihood is assigned automatically.

1.2.2 Continual Learning and Recyclable Inference

One of the most desirable properties for any modern machine learning method is the handling of very large datasets. Since this goal has been progressively achieved in the literature with scalable models, much attention is now paid to the notion of *efficiency*. For instance, in the way of accessing the data. The fundamental assumption uses to be that samples can be revisited without restrictions *a priori*. In practice, we encounter cases where the massive storage or data centralisation is not possible anymore for preserving the privacy of individuals, e.g. health and behavioral data. The mere limitation of data availability forces learning algorithms to derive new capabilities, such as i) distributing the data for *federated learning* (Smith et al., 2017), ii) observe streaming samples for *continual learning* (Goodfellow et al., 2014) and iii) limiting data exchange for *private-owned models* (Peterson et al., 2019).

A common theme in the previous approaches is the idea of model *memorising* and *recycling*, i.e. using the already fitted parameters in another problem or joining it with others for an additional global task without revisiting the collection of data. If we look to the functional view of this idea, uncertainty is still much harder to be repurposed than parameters. This is the point where GP models play they role.

Another contribution in this thesis is to provide a novel approach that extends the existing posterior-prior recursion of online Bayesian inference, e.g. with conjugate exponential-families, to the infinite-functional space setting of GP models. The key principle of the statistical method is the use of the conditional GP predictive equation to build *implicit* prior distributions where past posterior discoveries are propagated forward. We say the word *implicit* as long as the instances of inducing inputs are integrated, but the conditioning on their variational parameters remains. The entire model is amenable to stochastic optimization, letting us consider any irregular form in the sequential observation process. Additionally, the continual method is fully applicable to the multi-channel scenario, that is, to MOGP models as well as their heterogeneous counterpart.

Moreover, in this thesis we also investigate a general framework for recycling distributed variational approximations to GPs based on inducing inputs. Based on the fundamental properties of Kullback-Leibler divergence between stochastic processes (Matthews et al., 2016) and Bayesian inference, the proposed method ensembles an arbitrary amount of variational GP models with different complexity, likelihood function and location of pseudo-input instances, without revisiting any data. The recyclable framework is also amenable for regression, classification and heterogeneous tasks, and it is neither restricted to any specific sparse GP approach. We remark its potential use in applied cases, like large collections of patients data, where distributed databases are easily found.

1.2.3 Hierarchical Change-point Detection

Change-point detection (CPD), which consists of locating abrupt transitions in the generative model of the observations, is a problem with plethora of applications. In this thesis, we use CPD methods in behavioral data applied to mental health. However, the main focus of CPD has been traditionally on batch settings, where the entire sequence of observations is available and has to be segmented. This particular scenario is not well fitted to our application of interest. Instead, CPD is most useful for us in online settings, where change-points must be detected as new incoming samples are observed. This family of methods have two intertwined tasks to solve: i) segmentation of sequential data into partitions (or segments) and ii) estimation of the generative model parameters for the given partitions. Concretely, since each partition has a different generative distribution, the identifiability of the change-points is related to the difference between such densities, and hence, their parameters. In

this context, Bayesian inference is of particular utility for inferring the distributions given prior beliefs in a reliable manner (Adams and MacKay, 2007).

However, it can be observed that, for complex likelihood models, which have a growing number of parameters much higher than the amount of observations between two consecutive chage-points, reliable CPD becomes unfeasible. This use to be the case of, although is not restricted to, high-dimensional and/or heterogeneous observations, which usually have a prohibitive number of parameters for the CPD discrimination.

One of the main contributions of this thesis is to introduce a novel approach for CPD. To address the aforementioned issue, we present a hierarchical probabilistic model based on latent variables. The CPD problem can be carried out on the lower-dimensional manifold, where the discrete latent variables lie. The main advantage is that the method requires less evidence that the observational counterpart since the number of parameters is reduced. This yields faster and more reliable detections. However, the new statistical methods introduces new challenges that we also address for robust detection and its application to scenarios with an unbounded dimensionality of the latent structure. The key idea of this last model is to use the Chinese-restaurant process (CRP) (Pitman, 2002), which is a well-known Bayesian non-parametrics method to model the latent variables. Experimental results on real data show how the hierarchical approach performs reliably, which is a fundamental point for its application to human behavior problems.

1.2.4 Early Detection for Mental Health

For psychiatry, the principal advantage of personal mobile devices is that their embedded monitoring systems are completely unobtrusive for patients. This avoids short-term periodic interactions with clinicians, that are time-consuming for the healthcare system, and limits potential counfounders due to the self-representation. This one is sometimes the cause of subjective data gathering, i.e. via paper-and-pencil questionnaires, strongly dependent of the *static* behavioral state of the sufferer. Fortunately, these problems motivated the apparition of electronic mental health (e-Mental-Health) methods that nowadays are an emergent field (Osmani, 2015; Firth et al., 2016; Barrigón et al., 2017).

Regarding the comprehension of human behavioral dynamics and their digital phenotypes in affective disorders (Saha et al., 2016; Marzano et al., 2015), these novel pervasive services are ideally suited for capturing the behavioral states of patients during their daily routines. Existing approaches have already explored the analysis of human behavior from numerous modalities of information, such as activity recognition, communications registers, text and voice recognition, and similarly to the data considered in this thesis, mobility metrics.

Importantly, real-time prediction of triggering symptoms before an imminent relapse in mental health is strongly related to the detection of abrupt behavioral transitions. In this context, preventative medical interventions with automatic statistical methods in on the neccesity of using learning approaches where both dynamical characterization of profiles and detection of behavioral changes is performed as individuals emotional state may quickly vary during time.

The main applied contribution in this thesis is to propose a novel methodology for the detection of behavioral changes in psychiatric outpatients using different sources of daily information recorded from their personal smartphone. The key idea is to introduce probabilistic CPD methods for estimating the probabilities of change from the heterogeneous data of patients. In particular, we solve together both problems of heterogeneous likelihood functions and the high-dimensionality in CPD by assuming the hierarchical approach with latent variables. The experimental results in a preliminary study show the feasibility of the method for detecting behavioral changes in outpatients, and the easy intepretation for clinicians.

Finally, the CPD tool is put into practice within larger datasets for suicide attempt prevention in the context of the *Smartcrises* study protocol (Berrouiguet et al., 2019). Validation results within the clinical dates recorded from urgencies and hospitalary interventions show promising insights for the future of this technology in mental health.

1.3 Thesis Organization

The present doctoral manuscript is divided into four main chapters, that we shortly review in the following paragraphs for a better comprehension of the document and its organization. The references to all published pieces of work, where contributions were initially presented, are included in the introductory lines of each chapter.

Chapter 2: Models for Heterogeneous Data

This is perhaps, the first technical chapter of the thesis. Particularly, it focuses on the very first decisions to be taken for modelling both heterogeneous and irregular data observations, i.e. in presence of missing values. We begin with one simple approach based on discrete latent representations, such as latent class or mixture models, that accept multivariate likelihood distributions. Here, the features of observational vectors are accepted to belong to more-than-one statistical data type. This is, we mix binary and real-valued data in a same likelihood model with good results on inference. This same idea is later expanded to other types of probabilistic models, and particularly, to Gaussian processes (GP). The second half of the chapter is dedicated to the presentation of the heterogeneous multi-output GP model and its evaluation on different datasets with a purpose of scalability.

Chapter 3: Continual and Distributed Inference

In this chapter, we consider those scenarios where data cannot be accessed in a regular manner, that is, observations appear to be distributed or delivered to the system in a continual/online way. In particular, we propose several inference and learning methods based on stochastic processes, with special attention to GP models. In the first half, we present our idea for *recycling* already fitted models that we scale up via variational methods and GPs with arbitrary likelihood functions. In the next section, the previously described heterogeneous multi-output GP model in the preceding Ch. 2, is extended for being deployed in a continual learning setup. This methodology fuses concepts from sparse approximations to GPs, variational inference methods and signal processing.

Chapter 4: Change-point Detection

Once data modelling and inference methods are presented in Ch. 2 and Ch. 3 respectively, we focus on the problem of change-point detection (CPD). The fourth chapter develops a detailed analysis of methods for detecting abrupt distribution shifts in sequences of data with high-dimensionality. In addition, we also consider the rest of issues previously mentioned, e.g. heterogeneous likelihood functions and missing entries. Particularly, we introduce the hierarchical CPD model in the first half, which is an adaptation of Bayesian CPD to the presence of latent variables. This also allow us to scale up the model's order without restrictions, also improving its robustness. We propose several modifications of this idea, for instance, introducing the chinese-restaurant process (CRP) mechanism in the latent counterpart.

Chapter 5: Behavior Change Detection

All the technical advances and methods developed for data modelling, inference and change-point detection are put into practice in this chapter. The idea is to use behavioral models to detect change from smartphone-based data of psychiatric patients. The chapter is divided in three main sections. The first one presents the initial medical study where the probabilistic models were deployed for proving the viability of the hypothesis. In the second part, the clinical mental health study is extended for further applications, particularly for suicide attempt prevention. In this case, the multi-modal heterogeneous models presented in Ch. 2 are directly applied with a significant result. Finally, to demonstrate the significance and impact of the results obtained with both the behavior models and the change-point methods, detected events are compared and validated within clinical data from urgencies and hospital interventions.

Chapter 6: Conclusions and Future Work

We conclude the thesis by surveying the main technical contributions, as well as the advances based on the application of the human behavior models. As the reader should notice, this thesis combines ideas from multiple perspectives and research topics, to name a few, machine learning, behavior analysis, multivariate statistics and in the last case, mental health from a biomedical point of view, not fully clinical. Even being a first step towards the use of machine learning methods in behavioral related sciences and particularly in psychiatry, we provide additional ideas for the future development of both heterogeneous data models, distributed and continual inference and finally, change-point detection methods within latent variables.

CHAPTER 2

MODELS FOR HETEROGENEOUS DATA

PROBABILISTIC learning models are especially well-suited for the problem of *irregular* data observations. This sort of *complex* data can be easily found in any application with the purpose of modelling the human behavior. The main idea behind is that learning methods based on probability and uncertainty quantification are typically more robust to the apparition of input missings or unexpected events in the dataset. Among those models, such ones based on the concept of latent variables are of maximum interest for us. The ability to design low-dimensional latent manifolds with either continuous and discrete properties or even a temporal structure on demand is the first step forward to design reliable human behavior models for *irregular* observations.

In this chapter, we consider two different families of learning methods for the problem of human behavior modelling and its complementary heterogeneous data. Here, we refer to heterogeneous as the arbitrary combination of variables from different statistical data-types in the same set of observations, for example, a mix of binary, real-valued or categorical data.

We begin in Sec. 2.1 with the analysis of latent variable models in the context of low-dimensional discrete manifolds. We named them as latent *class* models following the original nomenclature used in Griffiths and Ghahramani (2011). As described in Sec. 2.1.2, we extended this family of latent variable models to accept heterogeneous likelihood distributions, and different mixes of statistical types are also considered. Additionally, a second extension is included to capture the periodic structure of the data via *circadian* covariance functions in Sec. 2.1.4. The same strategy is considered for a second family of latent variable models, Gaussian processes (GPs) (Rasmussen and Williams, 2006). In this case, latent variables are substituted with non-linear functions in the real-valued domain. In Sec. 2.2, we first introduce GP models from scratch followed by their adaptation to multi-task scenarios, well-known as multi-output GPs, in Sec. 2.2. Finally, in Sec. 2.2.1 we present a novel generalization of the GP models for handling heterogeneous likelihood functions, which is the main contribution of this chapter. In particular, this last model is capable to work with both large-scale data and irregular observations. The main details about the approximate inference carried out are described in Sec. 2.2.1 and further algorithms for parallel or continual extensions are included in the next Chap. 3.

The main technical advances described in this chapter have been previously presented in two research works. The first portions on heterogeneous latent class models and circadian covariance functions were included in Moreno-Muñoz et al. (2018, 2020), which was accepted in the Pattern Recognition Journal and it is pending for formal publication. Second, the results on heterogeneous GP models were presented in Moreno-Muñoz et al. (2018), that was accepted in the 2018 Conference on Advances in Neural Information Processing Systems (NeurIPS) as a *spotlight* oral presentation in Montréal, Canada.

2.1 Latent Variable Models for Heterogeneous Data

Unsupervised learning methods typically assume that there exists a low-dimensional latent structure responsible of the generative process of the data. In many well-known cases, this

structure refers to a common parameterization of the high-dimensional observations, for instance, given a linear mix of coefficients or real-valued factors that are later combined somehow. In others, a subset of auxiliary statistical variables are introduced for building the additive structure that helps in the problem of understanding the data. This ones are accepted to be both discrete or continuous, and depending on their nature, latent variable approaches are known with one name or another, e.g. factorial models or topic analysis. Additionally, one of the key problems in modern probabilistic ML is to face the decision on the nature of such latent structure and hence, its dimensionality. Often, practioners consider this as a problem of model selection.

In the last decades, the advances on latent variable models have been typically focused on the design of the latent structure and its variables, the corresponding parameterization and/or the inference mechanism required, which in many cases needs of approximations. However, a common ingredient among this family of probabilistic methods, is the connection of the hidden auxiliary objects with the high-dimensional observations via the likelihood distribution function. As a consequence, if both observations and latent variables lie under the assumption of independence and the distribution equality, then there are no restrictions for *heterogeneous* features in the data. Notice that this approach is more realistic as it enlarges the number of real-world scenarios where latent variable models could be applied. With a few exceptions (Khan et al., 2010; Valera et al., 2017), this idea has not been explored in depth. Consequently, we did a first exploration to it in this section.

2.1.1 Latent Class Models

One of the simplest types of latent variable models, is based on the idea that having observed N vectors $\boldsymbol{x}_i$, there exists a hidden *class* variable z_i which indicates the subset of generative properties of that particular vector. In this family of latent class models, such as *mixture* models (McLachlan and Basford, 1988; Bishop, 2006), the generative parameters are typically assigned to every class variable. Assuming that all the observations are in a larger vector $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \dots, \boldsymbol{x}_N^\top]^\top$, its latent counterpart is attached in the vector $\boldsymbol{z} = [z_1, z_2, \dots, z_N]^\top$. Typically, the joint probability distribution between these two vectors is given by the factorisation $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$, where the prior probability over the assignments $p(\boldsymbol{z})$ determines the nature of the model and initially sets the number of assignments K.

Importantly, the conditional probability distribution $p(\boldsymbol{x}|\boldsymbol{z})$, that we name here as the *likelihood* function of the latent variable model, is the one that determines how the class assignments z_i are related to the properties of observations $\boldsymbol{x}_i$. In this preliminary section, we analyse how this last distribution $p(\boldsymbol{x}|\boldsymbol{z})$ is the one that is capable to deal with multiple modalities, including heterogeneous features in a very simple manner. We remark that the traditional assumptions have been focused on the case of *homogeneous* data and the question of using *finite* or *infinite* orders for the dimensionality of the latent structure.

Finite Mixture Models

The well-known *mixture* models are a special case of latent *class* models, where the assignments of classes are independent among them. Particularly, the prior distribution $p(\boldsymbol{z})$ over the latent variables factorises according to the i.i.d. assumption. Thus, having a maximum of K class assignments, the distribution is

$$p(\boldsymbol{z}) = \prod_{i=1}^{N} p(z_i) = \prod_{i=1}^{N} \prod_{k=1}^{K} \pi_k^{\mathbb{I}\{z_i=k\}}, \tag{2.1}$$

• The i.i.d. assumption for class assignments is not true, for instance, in hidden markov models (HMM). In that case, assignments are conditioned to the previous one, given a 1st order Markov chain.

where we have considered that the latent variable marginals $p(z_i)$ are Categorical distributed. The expression $\mathbb{I}\{\cdot\}$ makes reference to the *indicator* function, whose output is *one* if the condition is satisfied and *zero* otherwise. The parameter $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_K]^\top$ is indexed under this factorisation by both the ith and kth values of z_i. Under this model, the conditional likelihood distribution of the observations takes the form

$$p(\boldsymbol{x}|\boldsymbol{z}) = \prod_{i=1}^{N}\prod_{k=1}^{K} p(\boldsymbol{x}_i|\boldsymbol{\theta}_k)^{\mathbb{I}\{z_i=k\}}, \tag{2.2}$$

where $p(\boldsymbol{x}_i|\boldsymbol{\theta}_k)$ is the likelihood probability of every observation $\boldsymbol{x}_i$ given the parameters $\boldsymbol{\theta}_k$ associated to the kth assignment of the latent class variables z_i. Importantly, the distribution $p(\boldsymbol{x}_i|\boldsymbol{\theta}_k)$ is often assumed to belong to one single family of probability density functions. For instance, when dealing with binary data, we typically place a Bernoulli mixture or Gaussian typed in the case of multivariate real-valued data (Bishop, 2006; Murphy, 2012). Looking to the properties of $p(\boldsymbol{x}_i|\boldsymbol{\theta}_k)$, we determine that it is particularly well suited for an extra factorization under the assumption of *conditional independence* (CI) in the dimensions of observations. The idea of introducing mixes of likelihood functions in the conditional term $p(\boldsymbol{x}_i|\boldsymbol{\theta}_k)$ makes us to consider this sort of latent class models. This is the first approach in the thesis for modelling irregular and heterogenous observations with probabilistic methods.

2.1.2 Heterogeneous Latent Class Models

Motivated by the application of probabilistic modelling to human behavior characterization, we consider data that posses some degree of heterogeneity and types of periodic temporal structure, with a 24h period, induced by the *circadian rhythm* of each individual. For this reason, we study here how to embed heterogeneous data models within the aforementioned temporal structure. The ideas presented in this section are also valid for any dataset with known periodicity.

To account for the periodic dependencies of the data, we propose to arrange them such that a single observed sample $\boldsymbol{x}_t$ at time t stacks the observations of one period as shown in Fig. 2.1. Additionally, we build heterogeneous observations by stacking dimensions from different statistical data-types on larger vectors $\boldsymbol{x}_t = [\boldsymbol{x}_t^1, \boldsymbol{x}_t^2, \ldots, \boldsymbol{x}_t^D]^\top$, with D being the number of heterogeneous data types. This means for us, that the sequence $\boldsymbol{x}_{1:t}$ summarizes consecutive instances (or time-steps) of different heterogeneous observations.

• Due to time is now considered in the data, we index objects as $\boldsymbol{x}_t$ instead of $\boldsymbol{x}_i$. Each sample $\boldsymbol{x}_t$ will now represent all the attributes recorded during 24h.

For the description of one single observation, represented by $\boldsymbol{x}_t$, we propose a heterogeneous latent class model. The likelihood distribution takes the form

$$p\left(\boldsymbol{x}_t|z_t, \{\boldsymbol{\theta}_k^1, \boldsymbol{\theta}_k^2, \ldots, \boldsymbol{\theta}_k^D\}_{k=1}^K\right) = \prod_{k=1}^{K}\prod_{d=1}^{D} p(\boldsymbol{x}_t^d|\boldsymbol{\theta}_k^d)^{\mathbb{I}\{z_t=k\}}, \tag{2.3}$$

where latent class variables z_t are the indicators of which component is active for the particular kth likelihood function. Concretely, the factorized expression in Eq. (2.3) is composed by K components, each one with its own likelihood distribution $p(\boldsymbol{x}_t^d|\boldsymbol{\theta}_k^d)$ for each data type. The variable $\boldsymbol{\theta}_k^d$ denotes the natural parameters of the dth data type that we aim to learn conditioned to the given latent class. It can also be seen that we have assumed *conditional independence* (CI) in Eq. (2.3), given the latent variable z_t and the parameters. This assumption simplifies the inference procedure, while at the same time makes the heterogeneous observations $\boldsymbol{x}_t^1, \boldsymbol{x}_t^2, \ldots, \boldsymbol{x}_t^D$ easy to be modelled.

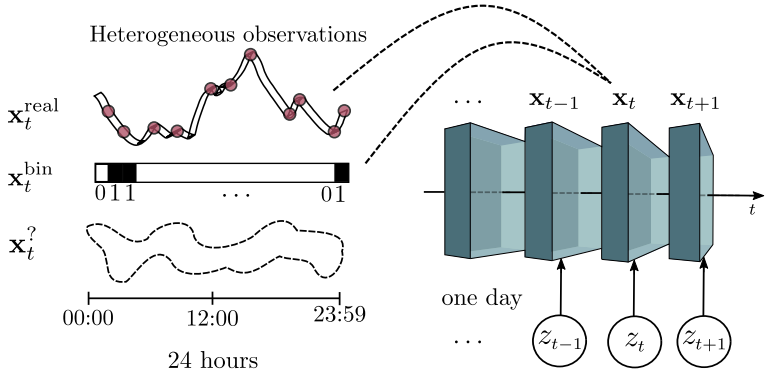


Figure 2.1: Schematic illustration of the embedded circadian model. Each green-shaded box represents a daily observation $\boldsymbol{x}_t$. Each observation is heterogeneous, i.e., mix of real, binary values, etc. Moreover, associated to $\boldsymbol{x}_t$, there exists a single latent variable z_t that is lower dimensional and we aim to discover.

2.1.3 Circadian Gaussian-Bernoulli Model

Among all heterogeneous statistical data types that may be incorporated to the daily representation $\boldsymbol{x}_t$, we choose here two representative cases for the considered application in human behavior learning. In particular, we assume that each observation $\boldsymbol{x}_t$ is composed by binary and real-valued features, that is, $\boldsymbol{x}_t = [(\boldsymbol{x}_t^{\text{real}})^\top, (\boldsymbol{x}_t^{\text{bin}})^\top]$, where the l.h.s. variables are $\boldsymbol{x}_t^{\text{real}} \in \mathbb{R}^p$ and $\boldsymbol{x}_t^{\text{bin}} = [x_{t1}^{\text{bin}}, \dots, x_{tp}^{\text{bin}}]^\top$, with $x_{tj}^{\text{bin}} \in \{0, 1\}$. Additionally, we use p to denote the dimensionality of each kind of observation, and in our particular case it is given by a function of the period. For example, if each component of $\boldsymbol{x}_t^{\text{real}}$ represents the travelled distance during a given hour, then $p = 24$, which corresponds to *one day*, the period induced by the circadian rhythm.

• We use letter p as the dimensionality of each vector attribute per data type, it refers to *precision*.

Note also that we are assuming that all data types have the same dimensionality p, but it would be straightforward to extend the model for different orders of dimensionality. Next, we select Bernoulli and Gaussian marginal distributions for the initial heterogeneous likelihood function, yielding

$$\begin{aligned} p(\boldsymbol{x}_t | z_t = k, \{\boldsymbol{\theta}_k\}_{k=1}^K) &= p(\boldsymbol{x}_t^{\text{bin}}, \boldsymbol{x}_t^{\text{real}} | \boldsymbol{\theta}_k) \\ &= \mathcal{N}(\boldsymbol{x}_t^{\text{real}} | \mathbf{0}, \mathbf{K}_k + \mathbf{D}) \prod_{j=1}^{p} \text{Ber}(x_{tj}^{\text{bin}} | \mu_{kj}), \end{aligned} \tag{2.4}$$

where, for the time being, we denote $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\gamma}_k\}$ as the set of likelihood parameters for each latent class value k. Here, $\boldsymbol{\mu}_k = [\mu_{k1}, \dots, \mu_{kp}]$, where each $\mu_{kj} \in [0, 1]$, is the mean of the jth Bernoulli feature and $\boldsymbol{\gamma}_k$ are the parameters of the covariance matrix $\mathbf{D}$. This one is chosen to be a common diagonal matrix for all latent classes z_t, and is given by $\mathbf{D} = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$, and the parameters of $\mathbf{K}_k$. In particular, the covariance matrices $\mathbf{K}_k$ are positive-definite and correspond to each class k as well. The purpose of the diagonal matrix $\mathbf{D}$ is to introduce *heteroscedasticity* in the model since we are assuming that the noise may vary within time inputs, that is, to be different across dimensions. A similar strategy will be later presented in Sec. 2.2.1 for Gaussian process models.

• Attributes in the 24h observation $\boldsymbol{x}_t$ are also correlated, i.e. morning actions condition night behavior.

To capture the periodic (circadian) feature of real-valued data $\boldsymbol{x}_t^{\text{real}}$, we employ a periodic covariance function, similar to the works in Solin and Särkkä (2014) and Durrande et al. (2016). Then, the matrix $\mathbf{K}_k$ is generated by a non-stationary periodic *kernel*, i.e.

$[\mathbf{K}_k]_{j,j'} = g_k(j, j')$, where the time or dimension j is assumed to be equally spaced (e.g. $j = 1, 2, 3, \ldots, p$). Further details about this function are provided in Sec. 2.1.4. Additionally, to obtain interesting insights, we consider that the distribution $\mathcal{N}(\boldsymbol{x}_t^{\text{real}}|\mathbf{0}, \mathbf{K}_k + \mathbf{D})$ is generated using a hierarchical methodology similar to the ones presented in Hensman et al. (2013b, 2015b). In our case, the mean vector $\boldsymbol{f}$ is drawn from a zero-mean Gaussian, such that $\boldsymbol{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_k)$, which yields $\boldsymbol{x}_t^{\text{real}}|\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{f}, \mathbf{D})$.

Following the standard approach used in latent class variable models, we need to compute the *complete* likelihood function, which included the prior probability for each latent variable indicator z_t. By defining the specific latent variable prior distribution as $p(z_t = k) = \pi_k$, the complete likelihood function at time t is

$$p(\boldsymbol{x}_t, z_t = k|\{\boldsymbol{\theta}_k\}_{k=1}^K) = p(z_t = k)p(\boldsymbol{x}_t|z_t = k, \boldsymbol{\theta}_k) = \pi_k p(\boldsymbol{x}_t|\boldsymbol{\theta}_k). \tag{2.5}$$

Next, taking into account that the heterogeneous observations are conditionally i.i.d., the complete joint log-likelihood finally becomes

$$\begin{aligned}\mathcal{L}_{\boldsymbol{z},\boldsymbol{\varphi}} = \log p(\boldsymbol{x}_{1:t}, \boldsymbol{z}_{1:t}|\{\boldsymbol{\theta}_k, \pi_k\}_{k=1}^K) = &\sum_{t=1}^{T}\sum_{k=1}^{K}\mathbb{I}\{z_t = k\}\log \pi_k \\ &+ \sum_{t=1}^{T}\sum_{k=1}^{K}\mathbb{I}\{z_t = k\}\log p(\boldsymbol{x}_t^{\text{bin}}|\boldsymbol{\mu}_k) \\ &+ \sum_{t=1}^{T}\sum_{k=1}^{K}\mathbb{I}\{z_t = k\}\log p(\boldsymbol{x}_t^{\text{real}}|\mathbf{K}_k, \mathbf{D}).\end{aligned} \tag{2.6}$$

One last comment is in order. If we need to take observation at non-uniformly spaced inputs j, say that the dimensions are not equally correlated as $j \neq 1, 2, 3, \ldots, p$ anymore, then the heterogeneous approach would also accept a different subset of inputs for each $\boldsymbol{x}_t$ in a straightforward manner. The reason behind is explained in the next section.

• We initially consider one heterogeneous attribute per hour, in the object $\boldsymbol{x}_t$. However, this regular sampling could be relaxed.

2.1.4 Circadian covariance functions

The proposed latent class model also captures the *circadian* feature of data for each kth assignment. In particular, it includes a temporal embedding and the use of periodic covariance functions $g_k(j, j')$, which in this case must be also non-stationary. Notice that afternoon-evening events should have a different correlation pattern than, for instance, nocturnal hours and early morning, which prevents the use of other stationary approximations. We propose to build non-stationary kernels with an input-dependent mapping $s_k(j)$, similarly to the one used in Heinonen et al. (2016). Here, we have $g_k(j, j') = s_k(j)s_k(j')\tilde{g}_k(j - j')$, where $\tilde{g}_k(j - j')$ is a stationary periodic covariance function, associated to the kth class. This models the intrinsic temporal structure during a day. In this case, we take the periodic version of the exponential kernel, which is given by

$$\tilde{g}_k(j - j') = \sigma_{ak}^2 \exp\left(-\frac{2\sin(\pi|j - j'|/D)}{\ell_k^2}\right), \tag{2.7}$$

and for $s_k(j)$, the hour-specific term, we use a squared Fourier series of order C, with $C \leq D$. This last constraint imposes a limit on the smoothness and avoids overfitting. Thus, $s_k(j)$ is

$$s_k(j) = \left(\frac{a_{k0}}{2} + \sum_{c=1}^{C}\left[a_{kc}\cos\left(\frac{2\pi c}{D}j\right) + b_{kc}\sin\left(\frac{2\pi c}{D}j\right)\right]\right)^2, \tag{2.8}$$

where $\mathbf{a}_k = [a_{k0}, \ldots, a_{kC}]^\top$ and $\mathbf{b}_k = [b_{k1}, \ldots, b_{kC}]^\top$ are the Fourier coefficients that parametrize the covariance matrix of the kth class, $\mathbf{K}_k$, together with the parameters of the exponential kernel, σ_{ak} and ℓ_k.

2.1.5 Alternative Latent Variable Models

There is plenty of extensive literature in the context of latent variable models for *homogeneous* datasets, where all the features that describe the high-dimensional observations are of the same statistical nature, either continuous or discrete. In the past decades, significant advances were done for mixed data scenarios, that is, pairs of one-to-one regression and classification problems (Khan et al., 2010). However, these models were often limited to the case of mixing two problems (regression and classification) at once and arbitrary combinations of other data types were not accepted.

Looking in depth to latent variable models, a few works considered this sort of approaches. We remark the importance of Khan et al. (2010), where factor analysis models are extended to accept a mix between continuous real-valued attributes and categorical observations. Importantly, this work shed light on the difficulties for performing inference given this sort of multivariate likelihood distributions, and variational methods had to be considered. Following a similar direction, Klami et al. (2012) used properties of exponential families to model heterogeneous coupled data (often named multi-view observations) via principal component analysis.

However, the recent alternative presented in Valera et al. (2017, 2020) opened the door to a more principled use of heterogeneous likelihood models in the context of latent variable models, in this case feature-based. The idea in this case is to assume that heterogeneous observations are mappings of some set of Gaussian distributed variables. This simple transformation simplifies significantly the number of assumptions to be taken, and at the same time, makes easier to accept other types of parameterizations and even latent variable structures. As a consequence of this advances, Nazabal et al. (2020) extended the idea for having a deep latent structure under an heterogeneous likelihood function to the domain of *variational autoencoders* (VAE). The balance between heterogeneous attributes has been recently improved in the context of deep generative models (Ma et al., 2020).

2.2 Gaussian Process Models for Heterogeneous Data

Probabilistic models based on Gaussian process (GP) methods (O'Hagan, 1992; Williams, 1998) are widely know for its flexible nature in the task of discovering correlations given input-output data observations. Since the interest caused by the formal presentation of GPs in Rasmussen and Williams (2006), 15 years ago, they have become a standard approach in the machine learning literature for modelling problems where non-linear parameterizations are needed, as well as strong uncertainty metrics for the posterior predicitive duties.

The main idea behind GP models is the specification of distributions over non-linear function spaces. This mechanism takes an important advantage. This is, the fact that the functions over such spaces are infinite dimensional objects. Thus, one can only instantiate the desired function values in several input points of interest if needed. However, the properties of the functional space still persist and may help the practitioner to perform any predictive task or conditional estimation. Despite there exist other several distributions to be considered for the infinite functional space (Shah et al., 2014), it is assummed to be Gaussian in this sort of stochastic processes. This decision is of particular relevance in the context of probabilistic modelling and particularly inference, as the integration becomes tractable in most cases.

In the very beginning, output observations were usually considered to be *noisy* versions of the function values, alike other regression methods in the literature. The noise model was assummed to be Gaussian and additive. This led most of the methods to link the underlying GP function to the mean parameter, e.g. first moment, of the real-valued data. In other words, just Gaussian likelihood distributions were considered for the output samples under study. In such cases, posterior inference only consisted of calculating the probabilities over the mean function, which coincides with the values of the underlying GP function. Additionally, it is well known that the assumption of a zero-mean prior in the functional space facilitates this same mechanism of inference, keeping it even simpler.

Putting all these ideas in the context of latent variables modelling, we can find a dual interpretation of the underlying (and unknown) function that parameterizes the likelihood distribution and hence, the data. In this case, the GP prior distribution plays a similar role as the marginal distribution on the latent variables that we introduced in Sec. 2.1.2. We also consider the key differences between both approaches, that in our present scenario, lead to not i.i.d. observations. Additionally, we see that there is still an open door in the formulation of GP models for introducing a similar methodology as the one considered with latent *class* models. This approach is the one that we develop in this section.

Formulation of Gaussian processes

Consider supervised learning scenarios where data consists of pairs of input-output observations $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^N$, with $\boldsymbol{x}_n \in \mathbb{R}^p$ and outputs y_n being either continous or discrete. From the perspective of a GP model, we assume that every output sample is generated as $y_n \sim p(y_n|f_n)$, where f_n is the non-linear function evaluation $f(\boldsymbol{x}_n)$. Here, the latent function $f(\cdot)$ that parameterizes the likelihood distribution is drawn from an infinite-dimensional GP prior, such that $f \sim \mathcal{GP}(0, k(\cdot,\cdot))$, where $k(\cdot,\cdot)$ can be any valid covariance function or *kernel*. The zero mean is assumed for simplicity in the inference.

Looking to one of its simpler versions, where we may consider Gaussian additive noise, the real-valued outputs y_n are assumed to be $y_n = f(\boldsymbol{x}_n) + \epsilon$. The noise is i.i.d. with variance σ_n^2. This is equivalent to say that the likelihood distribution over the vectors $\boldsymbol{y} = [y_1, y_2, \ldots, y_N]^\top$, takes the form $p(\boldsymbol{y}|\boldsymbol{f}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{f}, \sigma_n^2 I)$. Here, $\boldsymbol{f} = [f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \ldots, f(\boldsymbol{x}_N)]^\top$. The last term in the previous likelihood expression can also be rewritten as a zero-mean Gaussian distribution, $p(\boldsymbol{y}|0, \Sigma)$, where $\Sigma = k(\boldsymbol{x}, \boldsymbol{x}) + \sigma_n^2 I$. Notice that we stack input vectors to be $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]^\top$. Deriving the conditional distributions given the Gaussian likelihood model, we arrive to exact expressions for the posterior inference of $\boldsymbol{f}$ values, and hence, of the mean parameters for the output observations.

If we want to perform a predictive computation on some *test* input points $\boldsymbol{x}_*$, then we obtain the equations for Gaussian process regression (Rasmussen and Williams, 2006), that is

$$\boldsymbol{\mu}_* = k_*^\top (K + \sigma_n^2 I)^{-1} \boldsymbol{y}, \tag{2.9}$$

$$\boldsymbol{v}_* = k_{**} - k_*^\top (K + \sigma_n^2 I)^{-1} k_*, \tag{2.10}$$

where $k_* = k(\boldsymbol{x}_*, \boldsymbol{x})$, $k_{**} = k(\boldsymbol{x}_*, \boldsymbol{x}_*)$ and $K = k(\boldsymbol{x}, \boldsymbol{x})$. Both $\boldsymbol{\mu}_*$ and $\boldsymbol{v}_*$ represent the *mean* and *variance* of the estimation over the test points of interest.

The application of the predictive conditional equations in Eq. (2.9), Eq. (2.10), represents the main weakness of GP models. As the reader probably noticed, the inversion of K becomes undesirable as $N \to \infty$, and the number of operations turns to be unhandleable. Fortunately this problem is well known, and many works have been proposed since scaling up probabilistic machine learning models became a priority.

Sparse Approximations

Exact inference in GP models is widely known for its time complexity, which scales prohibitively to $\mathcal{O}(N^3)$ for inverting the kernel matrix and $\mathcal{O}(N^2)$ in memory cost. The computational constraints of standard implementations have motivated the development of many approximations in the literature. Among the principal works on scaling up GP models, we should remark those ones based on *sparse* methods (Seeger, 2003; Snelson and Ghahramani, 2006; Titsias, 2009a). The most part of this models construct approximations based on a smaller subset of $M \ll N$ *inducing* variables, which are also evaluations of the infinite-dimensional GP function. This approach allows to reduce the time complexity from $\mathcal{O}(N^3)$ to $\mathcal{O}(NM^2)$, as the true observations $\boldsymbol{y}$ are conditioned to such inducing variables that we aim to infer.

Usually, the difference between sparse methods rely on the way of treating the subset of inducing variables. A preliminary strategy in Snelson and Ghahramani (2006) considered such inducing variables as latent function values of f. These ones were later fitted as close as possible to the true output observations via maximum likelihood. Then, given a new input set $\boldsymbol{x}_*$, the predictive distribution of the GP is obtained as

$$\boldsymbol{\mu}_* = k_*^\top O_M^{-1} K_{MN} (\Lambda + \sigma_n^2 I)^{-1} \boldsymbol{y}, \tag{2.11}$$

$$\boldsymbol{v}_* = k_{**} - k_*^\top (K_M^{-1} - O_M^{-1}) k_* + \sigma_n^2, \tag{2.12}$$

where $O_M = K_M + K_{MN}(\Lambda + \sigma_n^2)_{-1} K_{MN}^\top$ and $\Lambda = \text{diag}(\boldsymbol{\lambda})$, with $\lambda_n = K_{nn} - k_n^\top K_M^{-1} k_n$. Notice that covariance matrices K_M and K_{NM} are built from the kernel function evaluations between the N input data points and the M inducing variables. The K_M is a square matrix. In terms of complexity, note that the computation of the predictive terms are dominated by the inversion of K_M, which is M long. Additionally, this inversion is later multiplied by the correlation matrix K_{NM}, what lead us to the final maximum cost of $\mathcal{O}(NM^2)$. This order of complexity has been maintained in the context of sparse approximations for GP models since its presentation, with a few recent exceptions (Wang et al., 2019) based on matrix decompositions. Beyond that, the original method in Snelson and Ghahramani (2006) still required to fit the pseudo-inputs in a non-smooth manner, and once put into practice, it was extremely difficult, e.g. find an optimal solution for the position the inducing variables close enough to the true N samples.

Based on the application of approximate marginal likelihood methods for fitting such *sparse* GP regression models, Titsias (2009a) introduced variational inference to estimate the location of inducing inputs and the adjacent hyperparameters. The key property of this new formulation was that it allowed to minimize the Kullback-Leibler (KL) divergence between the approximate variational distribution and the target posterior over the latent function values. In practice, to perform inference over the instances of the function f, Titsias (2009a) selected variational marginal densities of the form

$$q(\boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{f}|\boldsymbol{u}) q(\boldsymbol{u}), \tag{2.13}$$

where $\boldsymbol{u}$ are the function evaluations over the inducing inputs $\boldsymbol{z} = \{\boldsymbol{z}_m\}_{m=1}^M$ with $\boldsymbol{z}_m \in \mathbb{R}^p$, such that $\boldsymbol{u} = [f(\boldsymbol{z}_1), f(\boldsymbol{z}_2), \ldots, f(\boldsymbol{z}_M)]^\top$. The distribution $q(\cdot)$ is the density used to approximate the posterior at the collection of latent function values $\boldsymbol{u}$. The factorization in Eq. (2.13) will be recursively used in this thesis, particularly in the next Ch. 3.

The choice for $q(\boldsymbol{u})$ is typically a Gaussian distribution parameterized by the parameters $\boldsymbol{\mu}_{\boldsymbol{u}}$ and $\mathbf{S}_{\boldsymbol{u}}$, the mean vector and covariance matrix respectively. In the original formulation of Titsias (2009a), where an *homoscedastic* Gaussian likelihood model was chosen for the output observations, the lower bound under the log-marginal likelihood has a closed form.

Its expression is derived in Titsias (2009b). Additionally, this closed-form tractability leads to obtain unique optimal values for both $\boldsymbol{\mu}_{\boldsymbol{u}}$ and $\mathbf{S}_{\boldsymbol{u}}$. [1]

The approach presented in Titsias (2009a) was later extended in multiple directions. For the interest of this thesis, we remark three advances in the literature. First, the selection of the optimal number of inducing points M, which is the natural way to reduce the complexity of the approximate GP inference model. In addition to the perspective of Titsias (2009b), more recently, Burt et al. (2019, 2020) has investigated the convergence rates of inducing inputs in the context of sparse GP regression and the *vanilla* kernel. Likely, this advances will be expanded to other types of covariance functions and likelihood models. The idea is to know exactly *how many* inducing inputs are needed conditioned to the number of data N. Second, the additive Gaussian noise model was extended to accept *heteroscedastic* datasets. To do so, Lázaro-Gredilla and Titsias (2011) introduced a secondary GP function g to parameterize the likelihood variance, such that $\sigma_n^2 = \sigma^2(\boldsymbol{x}_n)\ \forall n \in \{1, 2, \ldots, N\}$, to be input-dependent. Finally, the original Gaussian likelihood model in Titsias (2009a) was substituted by others, for instance, Bernoulli distributions for scalable binary classification (Hensman et al., 2015a).

It is important to mention that the variational sparse methodology was also considered for multi-task or multi-output GP models, as we will see later in this section.

Multi-output Gaussian Processes

Multi-output Gaussian processes (MOGP), generalise the powerful Gaussian process (GP) predictive model to the vector-valued random field setup (Alvarez et al., 2012). Additionally, it has been experimentally shown that by simultaneously exploiting correlations between multiple outputs and across the input space, it is actually possible to provide better predictions, particularly in scenarios with missing or noisy data samples (Bonilla et al., 2008; Dai et al., 2017).

The main focus in the literature for GP and particularly MOGP models, has been on the definition of suitable cross-covariance functions between the multiple outputs that allow for the treatment of outputs as a single GP with a properly defined covariance function (Alvarez et al., 2012). The two classical alternatives to define such cross-covariance functions are the linear model of coregionalisation (LMC) (Journel and Huijbregts, 1978) and process convolutions (Higdon, 2002). In the former case, each output corresponds to a weighted sum of shared latent random functions. In the latter, each output is modelled as the convolution integral between a smoothing kernel and a latent random function common to all outputs. In both cases, the unknown latent functions follow Gaussian process priors leading to straightforward expressions to compute the cross-covariance functions among different outputs. More recent alternatives to build valid covariance functions for MOGP include the work by Ulrich et al. (2015) and Parra and Tobar (2017), that build the cross-covariances in the spectral domain.

Formulation of MOGP models

The use of Gaussian process models for modelling multiple-output observations, often named as *vector-valued* functions (Alvarez et al., 2012), follows a similar direction as in the single-output GP case. This type of probabilistic models, usually based on regression setups, has been largely study in the context of *geostatistical* scenarios, and in particular, the LMC approach.

The standard formulation for multi-output GPs considers a set of D dimensional outputs $\{y_d(\boldsymbol{x})\}_{d=1}^{D}$, with $y_d \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^p$. These real-valued outputs are also assummed to be

[1] See Eq. (10) in Titsias (2009a) for further details.

noisy versions of D latent output functions $\{f_d(\boldsymbol{x})\}_{d=1}^{D}$. In the case of LMC, each component f_d is assumed to be

$$f_d(\boldsymbol{x}) = \sum_{q=1}^{Q} a_{dq} u_q(\boldsymbol{x}), \tag{2.14}$$

where $u_q(\boldsymbol{x})$ are the *latent* functions and a_{dq} the mixing coefficients. At this point, we see the sort of approaches that this method introduces. The latent functions $u_q(\boldsymbol{x})$ can be modelled in very diffent manners, and in this thesis we will consider the use of *separate* GP priors. Thus, the processes $\{u_q(\boldsymbol{x})\}_{q=1}^{Q}$ are independent. This independence constraint can be also relaxed and even different mixings for the functions are allowed. The particular advantage of this structure is on the computation of cross-correlations between functions f_d, and hence, their corresponding output values y_d. An example of the exact computation of these cross-correlations will be later included in Sec. 2.2.1.

However, LMC approaches based on the *instantaneous* mixing of a subset of latent processes have some limitations in practice. For instance, it would be not possible to model multi-task problems with certain types of correlations, e.g. one output being a blurry version of another, or at list, the multi-output GP learning would be too limited. A powerful solution to these limitations is to introduce *convolution*. The idea of exploiting convolutions between base processes has been previously developed in Higdon (2002), extended in Wikle (2002) for spatiotemporal covariances and Paciorek and Schervish (2003) for non-stationary applications. The model introduced in Boyle and Frean (2004) considers that each f_d process is now expressed as a convolution integral of a latent process u_q and a smoothing kernel $G_d(\boldsymbol{x})$, such that

$$f_d(\boldsymbol{x}) = \int_{\mathcal{X}} G_d(\boldsymbol{x} - \boldsymbol{r}) u(\boldsymbol{r}) d\boldsymbol{r}. \tag{2.15}$$

Further details on the computation of cross-correlations between functions f_d and f'_d can be checked in Álvarez and Lawrence (2011). Importantly, even though convolutional processes (CP) in combination with multi-output GP models work better in practice, its computational complexity still scales up to $\mathcal{O}(Q^3N^3)$ for prediction and $\mathcal{O}(Q^3N^2)$ for storage. These costs are equivalent to the case of single-output GP regression but adapted to the Q dimensional multi-output setting. As a consequence, Alvarez and Lawrence (2009); Álvarez et al. (2010) introduced sparse approximations to the full covariance matrix, exploiting similar advantages as Titsias (2009a) did in the original scenario. The same approximations to multi-output GPs are applied in this thesis, but considering the simpler LMC model as a first approach for scaling up our methods. This is detailed in the following section.

2.2.1 Heterogeneous Multi-output Gaussian Processes

Regarding the statistical type of outputs that can be modelled by MOGP, most alternatives focus on multiple-output regression for continuous variables. Traditionally, each output is assumed to follow a Gaussian likelihood where the mean function is given by one of the outputs of the MOGP and the variance in that distribution is treated as an unknown parameter. Exact Bayesian inference is usually tractable for these models.

In this section, we are interested in the heterogeneous case for which the outputs are a mix of continuous, categorical, binary or discrete variables with different likelihood functions. A similar approach as the one we were intereseted in Sec. 2.1.2.

Consider a set of output functions $\mathcal{Y} = \{y_d(\boldsymbol{x})\}_{d=1}^{D}$, with $\boldsymbol{x} \in \mathbb{R}^p$, that we want to jointly model using Gaussian processes. Notice that each $y_d(\boldsymbol{x})$ is often assumed to be continuous and Gaussian distributed. Instead, we look to the scenario which outputs in $\mathcal{Y}$ are a mix

of continuous, categorical, binary or discrete variables with several generative distributions. In particular, we assume that the distribution over $y_d(\boldsymbol{x})$ is completely specified by a set of *input-dependent* parameters $\boldsymbol{\theta}_d(\boldsymbol{x}) \in \mathcal{X}^{J_d}$, where we have a generic $\mathcal{X}$ domain for the parameters and J_d is the number of parameters that define the distribution. Each parameter $\theta_{dj}(\boldsymbol{x}) \in \boldsymbol{\theta}_d(\boldsymbol{x})$ is a non-linear transformation of a Gaussian process prior function $f_{dj}(\boldsymbol{x})$, this is, $\theta_{dj}(\boldsymbol{x}) = g_{dj}(f_{dj}(\boldsymbol{x}))$, where $g_{dj}(\cdot)$ is a deterministic function that maps the GP output to the appropriate domain for the parameter θ_{dj}. To make this notation more concrete, let us assume an heterogeneous multi-output problem for which $D = 3$ as an example.

Bernoulli-Poisson-Gaussian Distributed Outputs

Assume that the output $y_1(\boldsymbol{x})$ is binary and that it is modelled using a Bernoulli distribution. The Bernoulli likelihood function uses a single parameter (the probability of success), $J_1 = 1$, restricted to values in the range $[0, 1]$. This means that $\boldsymbol{\theta}_1(\boldsymbol{x}) = \theta_{11}(\boldsymbol{x}) = g_{11}(f_{11}(\boldsymbol{x}))$, where $g_{11}(\cdot)$ could be modelled using the *logistic sigmoid* function $\sigma(\boldsymbol{z}) = 1/(1 + \exp(-\boldsymbol{z}))$ that maps $\sigma : \mathbb{R} \to [0, 1]$.

Second, assume that another output $y_2(\boldsymbol{x})$ corresponds to a counting variable that can take values $y_2(\boldsymbol{x}) \in \mathbb{N} \cup \{0\}$. This variable can be modelled using a Poisson distribution with a single parameter (the rate), $J_2 = 1$, restricted to the positive real numbers. This means that $\boldsymbol{\theta}_2(\boldsymbol{x}) = \theta_{21}(\boldsymbol{x}) = g_{21}(f_{21}(\boldsymbol{x}))$, where $g_{21}(\cdot)$ could be modelled as an exponential function $g_{21}(\cdot) = \exp(\cdot)$ to ensure strictly positive values for the parameter.

Finally, $y_3(\boldsymbol{x})$ could be a continuous variable with *heteroscedastic* noise. It can be modelled using a Gaussian distribution where both the mean and the variance are functions of $\boldsymbol{x}$. This means that $\boldsymbol{\theta}_3(\boldsymbol{x})$ would correspond to $\boldsymbol{\theta}_3 = [g_{31}(f_{31}(\boldsymbol{x})), g_{32}(f_{32}(\boldsymbol{x}))]^\top$, where the first function is used to model the mean of the Gaussian, and the second function is used to model the variance. Therefore, we can assume the $g_{31}(\cdot)$ is the identity function and $g_{32}(\cdot)$ is a mapping that ensures that the variance takes strictly positive values, e.g. the exponential function.

• Here, *heteroscedastic* makes reference to the property of input-dependent noise in the likelihood density. This leads to introduce a secondary GP function to parameterize the sigma parameter.

Heterogeneous Likelihood Formulation

Having the previous multivariate output data, let us now define a vector-valued function $\boldsymbol{y}(\boldsymbol{x}) = [y_1(\boldsymbol{x}), y_2(\boldsymbol{x}), \ldots, y_D(\boldsymbol{x})]^\top$. We assume that the outputs are conditionally independent given the vector of parameters $\boldsymbol{\theta}(\boldsymbol{x}) = [\theta_1(\boldsymbol{x}), \theta_2(\boldsymbol{x}), \ldots, \theta_D(\boldsymbol{x})]^\top$. These parametric vectors are defined by specifying the vector of latent functions

$$\boldsymbol{f}(\boldsymbol{x}) = [f_{11}(\boldsymbol{x}), f_{12}(\boldsymbol{x}), \ldots, f_{1J_1}(\boldsymbol{x}) f_{21}(\boldsymbol{x}), \ldots, f_{DJ_D}(\boldsymbol{x})]^\top \in \mathbb{R}^{J \times 1}, \tag{2.16}$$

where the total number of functions J is constrained to satisfy $J = \sum_{d=1}^{D} J_d$. The expression for the likelihood distribution is given by

$$p(\boldsymbol{y}(\boldsymbol{x})|\boldsymbol{\theta}(\boldsymbol{x})) = p(\boldsymbol{y}(\boldsymbol{x})|\boldsymbol{f}(\boldsymbol{x})) = \prod_{d=1}^{D} p(y_d(\boldsymbol{x})|\boldsymbol{\theta}_d(\boldsymbol{x})) = \prod_{d=1}^{D} p(y_d(\boldsymbol{x})|\tilde{\boldsymbol{f}}_d(\boldsymbol{x})), \tag{2.17}$$

where we have defined $\tilde{\boldsymbol{f}}_d(\boldsymbol{x}) = [f_{d1}(\boldsymbol{x}), f_{d2}(\boldsymbol{x}), \ldots, f_{dJ_D}(\boldsymbol{x})]^\top \in \mathbb{R}^{J_d \times 1}$, the set of latent functions that specify the parameters in $\boldsymbol{\theta}_d(\boldsymbol{x})$. Notice that $J \geq D$. This is, there is not always a one-to-one map from $\boldsymbol{f}(\boldsymbol{x})$ to $\boldsymbol{y}(\boldsymbol{x})$. The reader should notice that the factorization in Eq. (2.17) is analogous to the one used in the case of latent class models of Sec. 2.1.2. In this case, the latent functions f_d will be correlated together using the MOGP model.

Most previous work has assumed that $D = 1$, and that the corresponding elements in $\boldsymbol{\theta}_d(\boldsymbol{x})$, that is, the latent function values in $\tilde{\boldsymbol{f}}_1 = [f_{11}(\boldsymbol{x}), \ldots, f_{1J_1}(\boldsymbol{x})]^\top$ are drawn from

independent Gaussian processes. However, important exceptions on this point are Chai (2012) and Dezfouli and Bonilla (2015), that assumed a categorical variable $y_1(\boldsymbol{x})$, where the elements in $\tilde{\boldsymbol{f}}_1(\boldsymbol{x})$ were drawn from an intrinsic coregionalisation model.

In what follows in this section, we generalise the model for $D > 1$ and potentially heterogeneous outputs $y_d(\boldsymbol{x})$. The word "output" will be recursively used to refer to the elements $y_d(\boldsymbol{x})$ and "latent parameter function" (LPF) or "parameter function" (PF) to refer to $f_{dj}(\boldsymbol{x})$.

Multi-parameter GP priors

The main point of departure from previous work is in modelling function values $\boldsymbol{f}(\boldsymbol{x})$ using a multi-parameter Gaussian process prior that allows correlations for all the parameter functions $f_{dj}(\boldsymbol{x})$. The introduction of a linear model of corregionalisation type for covariance matrix functions facilitates the expresiveness of correlations between functions $f_{dj}(\boldsymbol{x})$ and $f_{d'j'}(\boldsymbol{x})$. The particular construction of these multi-parameter GP priors is as follows.

We consider an additional set of independent latent functions $\mathcal{U} = \{u_q(\boldsymbol{x})\}_{q=1}^{Q}$ that is linearly combined to produce J LPFs $\{f_{dj}(\boldsymbol{x})\}_{j=1,d=1}^{J_d,D}$. Each latent function or process $u_q(\boldsymbol{x})$ is assumed to be drawn from an independent GP prior such that $u_q \sim \mathcal{GP}(0, k_q(\cdot,\cdot))$, where k_q can be any valid covariance function, and the zero mean is assumed for simplicity. Then, each latent parameter $f_{dj}(\boldsymbol{x})$ is given as

$$f_{dj}(\boldsymbol{x}) = \sum_{q=1}^{Q}\sum_{i=1}^{R_q} a_{djq}^{i} u_q^{i}(\boldsymbol{x}), \tag{2.18}$$

where $u_q^i(\boldsymbol{x})$ are i.i.d. samples from $u_q(\cdot) \sim \mathcal{GP}(0, k_q(\cdot,\cdot))$ and $a_{djq}^i \in \mathbb{R}$. The mean function for $f_{dj}(\boldsymbol{x})$ is zero and the cross-covariance function $k_{f_{dj}f_{d'j'}}(\boldsymbol{x},\boldsymbol{x}') = \text{cov}[f_{dj}(\boldsymbol{x}), f_{d'j'}(\boldsymbol{x}')]$ is equal to $\sum_{q=1}^{Q} b^q_{(dj),(d'j')} k_q(\boldsymbol{x},\boldsymbol{x}')$, where $b^q_{(dj),(d'j')} = \sum_{i=1}^{R_q} a^i_{djq} a^i_{d'j'q}$.

The formal definition of the input data is given by $\boldsymbol{x} = \{\boldsymbol{x}_n\}_{n=1}^{N}$, and $\boldsymbol{x} \in \mathbb{R}^{N\times p}$, as a set of common vectors for all outputs $y_d(\boldsymbol{x})$. Although, the presentation of the model could be extended for the case of different set of inputs per output datum. We also define the vector-valued functions as $\boldsymbol{f}_{dj} = [f_{dj}(\boldsymbol{x}_1), f_{dj}(\boldsymbol{x}_2), \ldots, f_{dj}(\boldsymbol{x}_N)]^\top \in \mathbb{R}^{N\times 1}$; $\tilde{\boldsymbol{f}}_d = [\tilde{\boldsymbol{f}}_{d1}^\top, \ldots, \tilde{\boldsymbol{f}}_{dJ_d}^\top]^\top \in \mathbb{R}^{J_d N\times 1}$, and $\boldsymbol{f} = [\boldsymbol{f}_1^\top, \ldots, \boldsymbol{f}_D^\top]^\top \in \mathbb{R}^{JN\times 1}$.

Then, the generative model for the heterogeneous multi-output GP formulation is as follows. We sample output function values $\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{0}, \mathbf{K})$, where $\mathbf{K}$ is a block-wise matrix with blocks given by $\{\mathbf{K}_{\boldsymbol{f}_{dj}}, \mathbf{K}_{\boldsymbol{f}_{d'j'}}\}_{d=1,d'=1,j=1,j'=1}^{D,D,J_d,J_{d'}}$. In turn, the elements in $\mathbf{K}_{\boldsymbol{f}_{dj},\boldsymbol{f}_{d'j'}}$ are $k_{f_{dj},f_{d'j'}}(\boldsymbol{x}_n,\boldsymbol{x}_m)$, with $\boldsymbol{x}_n, \boldsymbol{x}_m \in \boldsymbol{x}$. For the particular case of equal input observations $\boldsymbol{x}$ for all LPFs, the matrix $\mathbf{K}$ can also be expressed as the sum of Kronecker products

$$\mathbf{K} = \sum_{q=1}^{Q} \mathbf{A}_q \mathbf{A}_q^\top \otimes \mathbf{K}_q = \sum_{q=1}^{Q} \mathbf{B}_q \otimes \mathbf{K}_q,$$

where $\mathbf{A}_q \in \mathbb{R}^{J\times R_q}$ has entries $\{a^i_{djq}\}_{d=1,j=1,i=1}^{D,J_d,R_q}$ and for $\mathbf{B}_q$ are $\{b^q_{(dj),(d'j')}\}_{d=1,d'=1,j=1,j'=1}^{D,D,J_d,J'_d}$. Thus, the matrix $\mathbf{K}_q \in \mathbb{R}^{N\times N}$ has entries given by $k_q(\boldsymbol{x}_n,\boldsymbol{x}_m)$ for $\boldsymbol{x}_n, \boldsymbol{x}_m \in \boldsymbol{x}$. Matrices $\mathbf{B}_q \in \mathbb{R}^{J\times J}$ are known as the *corregionalisation matrices*.

Once we are able to obtain vector-valued samples for $\boldsymbol{f}$, we may evaluate the proper vector of parameters $\boldsymbol{\theta} = [\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \ldots, \boldsymbol{\theta}_D^\top]^\top$, where we assumed that $\boldsymbol{\theta}_d = \tilde{\boldsymbol{f}}_d$. Having specified $\boldsymbol{\theta}$, we can generate samples for the output vector $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_D^\top]^\top \in \mathcal{X}^{DN\times 1}$, where the elements in $\boldsymbol{y}_d$ are obtained by sampling from the conditional distributions $p(y_d(\boldsymbol{x})|\boldsymbol{\theta}_d(\boldsymbol{x}))$.Importantly, to keep the notation uncluttered, we will assume from now that $R_q = 1$, meaning that

$\mathbf{A}_q = \mathbf{a}_q \in \mathbb{R}^{J\times 1}$, and the corregionalisation matrices are rank-one. In the literature, such model is widely known as the *semiparametric latent factor model* (Teh et al., 2005).

2.2.2 Scalable Variational Inference Framework

Given an heterogeneous dataset $\mathcal{D} = \{\boldsymbol{x}, \boldsymbol{y}\}$, we would like to compute the posterior distribution for $p(\boldsymbol{f}|\mathcal{D})$, which is intractable in our model. In what follows, we use similar ideas to Alvarez and Lawrence (2009); Álvarez et al. (2010) that introduced the inducing variable formalism of Sec. 2.2 for computational efficiency in MOGP models. However, instead of explicitly marginalising the underlying latent functions $\mathcal{U}$ to obtain a variational lower bound, we keep their presence in a way that allows us to apply stochastic variational inference as in Hensman et al. (2013a); Saul et al. (2016).

Inducing Variables for Heterogeneous MOGP

A key idea to reduce the computational complexity of Gaussian process models is to introduce *auxiliary variables* or *inducing variables*, which typically open the door to *sparse approximations* (Sec. 2.2). The variables have been used already in the context of MOGP models (Álvarez et al., 2009, 2010). A subtle difference from the single output case in Sec. 2.2 is that the inducing variables are no longer taken from the same latent process, say $f(\boldsymbol{x})$, but from the latent processes $\mathcal{U}$ instead. Those are used also to build the model for multiple-outputs. We follow the same formalism here for heterogeneous data.

We start by defining the set of M inducing variables per latent function $u_q(\boldsymbol{x})$ as $\boldsymbol{u} = [u_q(\boldsymbol{z}_1), \dots, u_q(\boldsymbol{z}_M)]^\top$, evaluated at a set of inducing inputs $\boldsymbol{z} = \{\boldsymbol{z}_m\}_{m=1}^M \in \mathbb{R}^{M\times p}$. We also define the vector $\boldsymbol{u} = [\boldsymbol{u}_1^\top, \dots, \boldsymbol{u}_Q^\top]^\top \in \mathbb{R}^{QM\times 1}$. For simplicity in the exposition, the general assumption is that all the inducing variables, for all q, are evaluated at the same subset of *pseudo*-inputs $\boldsymbol{z}$.

Instead of marginalising $\{u_q(\boldsymbol{x})\}_{q=1}^Q$ from the standard LMC model in Eq. (2.18), we explicitly use the joint Gaussian prior $p(\boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u})$. Due to the assumed independence in the latent functions $u_q(\boldsymbol{x})$, the distribution $p(\boldsymbol{u})$ factorises according to $p(\boldsymbol{u}) = \prod_{q=1}^Q p(\boldsymbol{u}_q)$, with $\boldsymbol{u}_q \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_q)$. Here, matrices $\mathbf{K}_q \in \mathbb{R}^{M\times M}$ has entries $k_q(\boldsymbol{z}_i, \boldsymbol{z}_j)$ with $\boldsymbol{z}_i, \boldsymbol{z}_j \in \boldsymbol{z}$. Notice that in this case, the dimensions of matrices $\mathbf{K}_q$ are different to the ones in the previous Sec. 2.2.1, since we now consider the *sparse* approximation.

The LPFs instances $\boldsymbol{f}_{dj}$ are conditionally independent given $\boldsymbol{u}$, so we can rewrite the conditional distribution $p(\boldsymbol{f}|\boldsymbol{u})$ as

$$
\begin{aligned}
p(\boldsymbol{f}|\boldsymbol{u}) &= \prod_{d=1}^{D}\prod_{j=1}^{J_d} p(\boldsymbol{f}_{dj}|\boldsymbol{u}) \\
&= \prod_{d=1}^{D}\prod_{j=1}^{J_d} \mathcal{N}\left(\boldsymbol{f}_{dj}|\mathbf{K}_{\boldsymbol{f}_{dj}\boldsymbol{u}}\mathbf{K}_{\boldsymbol{u}\boldsymbol{u}}^{-1}\boldsymbol{u}, \mathbf{K}_{\boldsymbol{f}_{dj}\boldsymbol{f}_{dj}} - \mathbf{K}_{\boldsymbol{f}_{dj}\boldsymbol{u}}\mathbf{K}_{\boldsymbol{u}\boldsymbol{u}}^{-1}\mathbf{K}_{\boldsymbol{f}_{dj}\boldsymbol{u}}^\top\right),
\end{aligned} \tag{2.19}
$$

where $\mathbf{K}_{\boldsymbol{u}\boldsymbol{u}} \in \mathbb{R}^{QM\times QM}$ is a block-diagonal matrix with sub-blocks given by $\mathbf{K}_q$ and $\mathbf{K}_{\boldsymbol{f}_{dj}\boldsymbol{u}} \in \mathbb{R}^{N\times QM}$ is the cross-covariance matrix computed from the cross-covariances between $f_{dj}(\boldsymbol{x})$ and $u_q(\boldsymbol{z})$. The expression for this cross-covariance function can be obtained from Eq. (2.18), leading to $k_{f_{dj}u_q}(\boldsymbol{x}, \boldsymbol{z}) = a_{djq}k_q(\boldsymbol{x}, \boldsymbol{z})$. More details can be also revisited in Alvarez et al. (2012). This form for the cross-covariance between the LPF $f_{dj}(\boldsymbol{x})$ and $u_q(\boldsymbol{z})$ is a key difference between the inducing variables methods for the single-output GP case and the MOGP case.

Construction of Variational Bounds

As in other GP methods with non-Gaussian likelihood functions, exact posterior inference is intractable. However our priority is to model the heterogeneous data in an scalable manner. For this task, we use variational inference methods to compute a lower bound $\mathcal{L}$ for the marginal log-likelihood $\log p(\boldsymbol{y})$, and for approximating the target posterior distribution $p(\boldsymbol{f}, \boldsymbol{u}|\mathcal{D})$. We follow the notation of Álvarez et al. (2010), where the posterior distribution over the LPFs $\boldsymbol{f}$ and the LFs $\boldsymbol{u}$ can be approximated via

$$p(\boldsymbol{f}, \boldsymbol{u}|\boldsymbol{y}, \boldsymbol{x}) \approx q(\boldsymbol{f}, \boldsymbol{u}) = p(\boldsymbol{f}|\boldsymbol{u})q(\boldsymbol{u}) = \prod_{d=1}^{D}\prod_{j=1}^{J_d} p(\boldsymbol{f}_{dj}|\boldsymbol{u}) \prod_{q=1}^{Q} q(\boldsymbol{u}_q), \tag{2.20}$$

where $q(\boldsymbol{u}_q) = \mathcal{N}(\boldsymbol{u}_q|\boldsymbol{\mu}_{\boldsymbol{u}_q}, \mathbf{S}_{\boldsymbol{u}_q})$ are the auxiliary variational distributions whose natural parameters $\{\boldsymbol{\mu}_{\boldsymbol{u}_q}, \mathbf{S}_{\boldsymbol{u}_q}\}_{q=1}^{Q}$ we aim to optimize for maximising our lower bound $\mathcal{L}$. Building on previous work by Saul et al. (2016), we derive a lower bound that accepts any log-likelihood function that can be modulated by the LPFs $\boldsymbol{f}$. The lower bound $\mathcal{L}$ for $\log p(\boldsymbol{y})$ can be obtained as follows

$$\begin{aligned}\log p(\boldsymbol{y}) &= \log \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u})d\boldsymbol{f}d\boldsymbol{u} \\ &\geq \int q(\boldsymbol{f}, \boldsymbol{u}) \log\left(\frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u})}{q(\boldsymbol{f}, \boldsymbol{u})}\right) d\boldsymbol{f}d\boldsymbol{u} = \mathcal{L}.\end{aligned} \tag{2.21}$$

We can further simplify the previous expression of $\mathcal{L}$ in Eq. (2.21) to obtain

$$\begin{aligned}\mathcal{L} &= \iint p(\boldsymbol{f}|\boldsymbol{u})q(\boldsymbol{u}) \log p(\boldsymbol{y}|\boldsymbol{f})d\boldsymbol{f}d\boldsymbol{u} - \sum_{q=1}^{Q} \mathrm{KL}\left[q(\boldsymbol{u}_q)||p(\boldsymbol{u}_q)\right] \\ &= \iint \prod_{d=1}^{D}\prod_{j=1}^{J_d} p(\boldsymbol{f}_{dj}|\boldsymbol{u})q(\boldsymbol{u}) \log p(\boldsymbol{y}|\boldsymbol{f})d\boldsymbol{u}d\boldsymbol{f} - \sum_{q=1}^{Q} \mathrm{KL}\left[q(\boldsymbol{u}_q)||p(\boldsymbol{u}_q)\right],\end{aligned} \tag{2.22}$$

where KL is the Kullback-Leibler divergence. Moreover, the approximate marginal posterior for $\boldsymbol{f}_{dj}$ is $q(\boldsymbol{f}_{dj}) = \int p(\boldsymbol{f}_{dj}|\boldsymbol{u})q(\boldsymbol{u})d\boldsymbol{u}$, leading to the exact expression

$$q(\boldsymbol{f}_{dj}) = \mathcal{N}\left(\boldsymbol{f}_{dj}|\mathbf{K}_{\boldsymbol{f}_{dj}\boldsymbol{u}}\mathbf{K}_{\boldsymbol{u}\boldsymbol{u}}^{-1}\boldsymbol{\mu}_{\boldsymbol{u}}, \mathbf{K}_{\boldsymbol{f}_{dj}\boldsymbol{f}_{dj}} + \mathbf{K}_{\boldsymbol{f}_{dj}\boldsymbol{u}}\mathbf{K}_{\boldsymbol{u}\boldsymbol{u}}^{-1}\left(\mathbf{S}_{\boldsymbol{u}} - \mathbf{K}_{\boldsymbol{u}\boldsymbol{u}}\right)\mathbf{K}_{\boldsymbol{u}\boldsymbol{u}}^{-1}\mathbf{K}_{\boldsymbol{f}_{dj}\boldsymbol{u}}^{\top}\right), \tag{2.23}$$

where $\boldsymbol{\mu}_{\boldsymbol{u}} = [\boldsymbol{\mu}_{\boldsymbol{u}_1}^{\top}, \ldots, \boldsymbol{\mu}_{\boldsymbol{u}_Q}^{\top}]^{\top}$ and $\mathbf{S}_{\boldsymbol{u}}$ is a block-diagonal matrix with blocks given by the variational matrices $\mathbf{S}_{\boldsymbol{u}_q}$. Importantly, the log-conditional distribution $\log p(\boldsymbol{y}|\boldsymbol{f})$, that here we know as the likelihood function of the heterogeneous model, factorises according to the previous Eq. (2.17) as

$$\log p(\boldsymbol{y}|\boldsymbol{f}) = \sum_{d=1}^{D} \log p(\boldsymbol{y}_d|\tilde{\boldsymbol{f}}_d) = \sum_{d=1}^{D} \log p(\boldsymbol{y}_d|\boldsymbol{f}_{d1}, \boldsymbol{f}_{d2}, \ldots, \boldsymbol{f}_{dJ_d}). \tag{2.24}$$

Using the expression for $\log p(\boldsymbol{y}|\boldsymbol{f})$, we can rewrite the expression for the variational bound $\mathcal{L}$ as

$$\mathcal{L} = \sum_{d=1}^{D} \mathbb{E}_{q(\boldsymbol{f}_{d1})\ldots q(\boldsymbol{f}_{dJ_d})}\left[\log p(\boldsymbol{y}_d|\boldsymbol{f}_{d1}, \ldots, \boldsymbol{f}_{dJ_d})\right] - \sum_{q=1}^{Q} \mathrm{KL}\left[q(\boldsymbol{u}_q)||p(\boldsymbol{u}_q)\right]. \tag{2.25}$$

When $D = 1$ in the expression above, we recover the same bound as in Saul et al. (2016). To maximise this lower bound $\mathcal{L}$, we need to find the optimal variational parameters $\{\boldsymbol{\mu}_{\boldsymbol{u}_q}\}_{q=1}^{Q}$ and $\{\mathbf{S}_{\boldsymbol{u}_q}\}_{q=1}^{Q}$. In this particular model, we choose to represent each matrix $\mathbf{S}_{\boldsymbol{u}_q}$ as the product of lower-triangular matrices $\mathbf{S}_{\boldsymbol{u}_q} = \mathbf{L}_{\boldsymbol{u}_q}\mathbf{L}_{\boldsymbol{u}_q}^{\top}$ and, to ensure the positive definiteness for $\mathbf{S}_{\boldsymbol{u}_q}$, we estimate $\mathbf{L}_{\boldsymbol{u}_q}$ instead of $\mathbf{S}_{\boldsymbol{u}_q}$.

Approximate Methods for Variational Expectations

There are still intractable issues in the variational expectations (or integrals) on the log-likelihood functions. Since we construct these bounds in order to acccept any possible statistical data-type, we need a general way to solve this sort of integrals. One obvious solution is to apply Monte Carlo (MC) methods, however it would be too slow both maximising the lower bound and updating variational parameters by sampling thousands of times (for approximating expectations) at each iteration of the optimization process. Instead, we address this problem by using Gaussian-Hermite (GH) quadratures as in Hensman et al. (2015a) and Saul et al. (2016).

One example with GH quadratures is in order. To compute the integrals of the form

$$\mathbb{E}_{q(\tilde{\boldsymbol{f}}_d)}\big[\log p(\boldsymbol{y}_d|\tilde{\boldsymbol{f}}_d)\big] = \int q(\tilde{\boldsymbol{f}}_d)\log p(\boldsymbol{y}_d|\tilde{\boldsymbol{f}}_d)d\tilde{\boldsymbol{f}}_d, \tag{2.26}$$

that we here consider to be univariate for simplicity, we apply the approximation

$$\mathbb{E}_{q(\boldsymbol{f}_{d1})}\big[\log p(\boldsymbol{y}_d|\boldsymbol{f}_{d1})\big] \approx \frac{1}{\sqrt{\pi}}\sum_{s=1}^{S} w_s \log p(\boldsymbol{y}_d|\sqrt{2\boldsymbol{v}_{d1}}\boldsymbol{f}_s + \boldsymbol{m}_{d1}), \tag{2.27}$$

where $\boldsymbol{m}_{d1}$ and $\boldsymbol{v}_{d1}$ are the mean and variance of the variational distribution $q(\boldsymbol{f}_{d1})$, respectively. In addition, the pair of values $\{w_s, \boldsymbol{f}_s\}$ is obtained by taking a chosen number S of points from the Hermite polynomial

$$H_n(x) = (-1)^n e^{x^2}\frac{d^n}{dx^n}e^{-x^2}.$$

Importantly, notice that this process must be done sequentially for multivariate expectations, which results in a multidimensional sum with an storage cost of $\mathcal{O}(S^{D_t})$ where D_t is the number of output functions involved in the integral. For very large-dimensional Categorical output, new ways of parameterizations should be considered, for instance, Ruiz et al. (2018).

Stochastic Variational Inference

Turning back to the formulation of the lower bound, the conditional expectations in Eq. 2.25 above are also valid across data observations, so that we can express the bound as

$$\mathcal{L} = \sum_{d=1}^{D}\sum_{n=1}^{N}\mathbb{E}_{q(f_{d1}(\boldsymbol{x}_n)),\ldots,q(f_{dJ_d}(\boldsymbol{x}_n))}\left[\log p(y_d(\boldsymbol{x}_n)|f_{d1}(\boldsymbol{x}_n),\ldots,f_{dJ_d}(\boldsymbol{x}_n))\right] - \sum_{q=1}^{Q}\mathrm{KL}\left[q(\boldsymbol{u}_q)||p(\boldsymbol{u}_q)\right]. \tag{2.28}$$

This functional form allows the use of *mini-batches* of smaller sets of training samples, performing the optimization process using noisy estimates of the global objective gradient (Robbins and Monro, 1951). Similar strategies have been used in Hoffman et al. (2013); Hensman et al. (2013b, 2015a) and Saul et al. (2016).

An scalable bound of this type makes our multi-output model applicable to large heterogeneous datasets. Importantly, we remark that the computational complexity is dominated by the inversion of $\mathbf{K}_{\boldsymbol{uu}}$ with a cost of $\mathcal{O}(QM^3)$ and products like $\mathbf{K}_{\boldsymbol{fu}}$ with a typical cost of $\mathcal{O}(JNQM^2)$. However, several extensions can be adopted in the parameterization of the likelihood functions, so these costs could be even smaller.

Learning of Hyperparameters and Prediction

The hyperparameters of the heterogeneos MOGP include $\boldsymbol{z}, \{\mathbf{B}_q\}_{q=1}^{Q}$, and $\{\boldsymbol{\gamma}_q\}_{q=1}^{Q}$, which are the hyperparameters associated to the covariance functions $\{k_q(\cdot,\cdot)\}_{q=1}^{Q}$. Since the variational distribution $q(\boldsymbol{u})$ is sensitive to changes of the hyperparameters, we maximise the variational parameters for $q(\boldsymbol{u})$, and the hyperparameters using a variational EM algorithm (Beal, 2003) when employing the full dataset, or the stochastic version when using mini-batches (Hoffman et al., 2013).

Moreover, for the computation of predictive probabilities, we consider the set of test inputs $\boldsymbol{x}_*$. Assuming that $p(\boldsymbol{u}|\boldsymbol{y}) \approx q(\boldsymbol{u})$, the predictive distribution $p(\boldsymbol{y}_*|\boldsymbol{y})$ associated to $\boldsymbol{x}_*$, can be approximated as $p(\boldsymbol{y}_*|\boldsymbol{y}) \approx \int p(\boldsymbol{y}_*|\boldsymbol{f}_*)q(\boldsymbol{f}_*)d\boldsymbol{f}_*$, where at the same time $q(\boldsymbol{f}_*) = \int p(\boldsymbol{f}_*|\boldsymbol{u})q(\boldsymbol{u})d\boldsymbol{u}$. The predictive variational expression factorizes according to

$$q(\boldsymbol{f}_*) = \prod_{d=1}^{D}\prod_{j=1}^{J_d} q(\boldsymbol{f}_{dj*}), \tag{2.29}$$

and it involves evaluating matrices $\mathbf{K}_{\boldsymbol{f}_{dj*}\boldsymbol{u}}$ at $\boldsymbol{x}_*$. As in the case of the expectation integrals in the lower bound of Eq. (2.25), the integral for $p(\boldsymbol{y}_*|\boldsymbol{y})$ above is intractable for the non-Gaussian likelihoods $p(\boldsymbol{y}_*|\boldsymbol{f}_*)$. For this integrals, we can make use of MC methods or quadratures to approximate them. Simpler integration problems are obtained if we are only interested in the predictive mean, $\mathbb{E}[\boldsymbol{y}_*|\boldsymbol{y}]$, and the predictive variance, $\text{var}[\boldsymbol{y}_*|\boldsymbol{y}]$. The expressions of both predictive terms are given by

$$\mathbb{E}\big[\boldsymbol{y}_d^*|\boldsymbol{y}\big] = \int \mathbb{E}\big[\boldsymbol{y}_d^*|\tilde{\boldsymbol{f}}_d^*\big]q(\tilde{\boldsymbol{f}}_d^*)d\tilde{\boldsymbol{f}}_d^*, \tag{2.30}$$

$$\text{var}\big[\boldsymbol{y}_d^*|\boldsymbol{y}\big] = \int \text{var}\big[\boldsymbol{y}_d^*|\tilde{\boldsymbol{f}}_d^*\big]q(\tilde{\boldsymbol{f}}_d^*)d\tilde{\boldsymbol{f}}_d^* + \int \mathbb{E}\big[\boldsymbol{y}_d^*|\tilde{\boldsymbol{f}}_d^*\big]^2 q(\tilde{\boldsymbol{f}}_d^*)d\tilde{\boldsymbol{f}}_d^* - \mathbb{E}\big[\boldsymbol{y}_d^*|\boldsymbol{y}\big]^2. \tag{2.31}$$

Note that prediction can be performed independently for each dth output type y_d given their corresponding LPFs f_d.

2.2.3 Heterogeneous Single-output Gaussian Processes

It is important to mention that the we have considered so far heterogeneous data problems in a vector-valued manner. In both Sec. 2.1.2 and Sec. 2.2.1, target observations are modelled as multimodal vectors, where each dimension belongs to a different statistical data type, e.g. the nth output of the heterogeneous MOGP is assumed to be $\boldsymbol{y}_n = [y_1(\boldsymbol{x}_n), y_2(\boldsymbol{x}_n), \ldots, y_D(\boldsymbol{x}_n)]^\top$. However, this assumption can be relaxed to obtain more flexible models where each univariate output observation y_n is of different nature (either continuous or discrete) and we model it using a single-output GP prior.

This is a natural extension of GPs to heterogeneous likelihood models with a slightly different structure than the one used in Sec. 2.2.1 and in Moreno-Muñoz et al. (2018). In that cases, there are no restrictions to accept a single-output GP function, such that densities $p(y_n|f(\boldsymbol{x}_n))$ may change per data point $\{\boldsymbol{x}_n, y_n\}$. The only difference with the key stochastic variational approximations in the literature (Hensman et al., 2013a, 2015a) is that the

expectation integrals in the lower bound (similarly as in Eq. (2.28)) may also be computed over several likelihood functions instead, one per data point. However, we identify some inconvenients in this type of modelling. Each ith expectation term could be *imbalanced* with respect to the others. For instance, if mixing Bernoulli and Gaussian distributed variables, binary outputs could contribute more to the objective function than the rest, due to the dimensionality. To overcome this issue, local GP models could be considered as in Moreno-Muñoz et al. (2020a). This will be analyzed, from a technical point of view, later in Ch. 3. Another drawback of this approach is that data-types need to be known beforehand, perhaps as additional *labels*.

• The term *imbalance* here refers to the magnitudes of probability values. Mixing a p.m.f. with a p.d.f. for instance, could be problematic.

2.3 Evaluation of Models for Heterogeneous Data

In this section, we aim to evaluate the performance of the latent variable model presented in Sec. 2.1.2 and the Gaussian process model of Sec. 2.2.1 for heterogeneous data also developed in this chapter. The current section is therefore divided into three main blocks. First, the experimental results presented in the next Sec. 2.3.1 correspond to the *circadian* Gaussian-Bernoulli model of the previous Sec. 2.1.3. Its evaluation is divided between the Ch. 2 and Ch. 4, within the change-point detection approaches. Here, we focus exclusively on the ability of the model for capturing the underlying periodic features from heterogeneous distributed data. Second, another set of experiments is dedicated to the analysis of heterogeneous MOGPs, that we formulated in Sec. 2.2.1. The empirical validation is performed accross different datasets (both synthetic and real-world) oriented to the general application of this thesis. We pay special attention to the modelling of demographic and human behavior data collections as well as we aim to demonstrate its scalability for large scale datasets. Finally, we describe extra simulations in the context of heterogeneous single-output GPs.

2.3.1 Heterogeneous Latent Class Model Simulations

In this experiment, we considered data that consists of location raw traces (latitude-longitude coordinates) recorded via the personal smartphone of a student of our laboratory. The observations were recorded during 275 consecutive days. The collection contains a bit more than 100K instances that correspond to the user's GPS coordinates every 3 minutes on average. In this case, we considered two types of metrics (Canzian and Musolesi, 2015): i) a real-valued signal of the log-distance travelled *per hour* and ii) daily binary vectors of *presence* or *absence* at home. Further details on these data are included in Ch. 5, where the applications to human behavior are reviewed.

Circadian Phenotypes

One of the strengths of the heterogeneous latent class model described in Sec. 2.1.2 is the ability to capture the 24h periodic structure from the collection of mobility metrics. It is based on the estimation of Fourier coefficients for the covariance function given the distance values. After fitting all hyperparameters in the maximization step during the EM inference algorithm, we are able to reproduce the patterns that describe how a person usually behaves during each type of day. In Fig. 2.2, we represent the set of multimodal periodic functions generated from the squared Fouries series, $s_k(j)$, and the estimated vectors of probability $\boldsymbol{\mu}_k$ of the Bernoulli likelihood density, given the heterogeneous model.

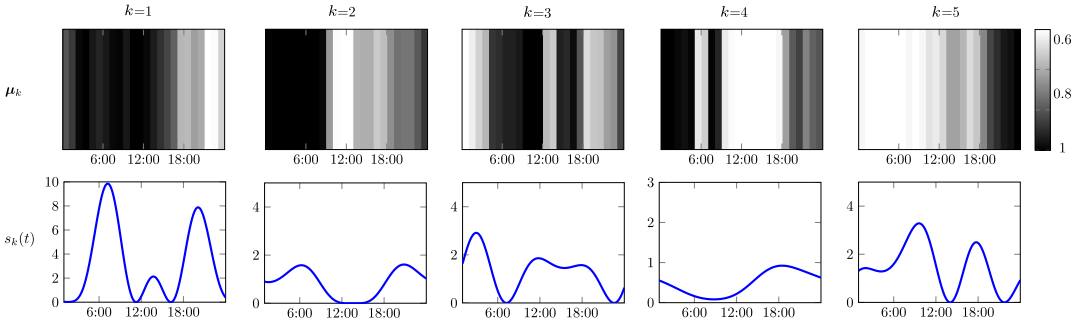


Figure 2.2: Circadian patterns of mobility. The results are obtained for $K = 5$. (UPPER ROW) Periodic functions generated from the squared Fourier series with the estimated hyperparameters. Functions represent the variance in the mobility per hour and type of day. (LOWER ROW) Thermal plots of multivariate Bernoulli distribution, i.e., the probability of being at-home during different hours.

2.3.2 Heterogeneous Multi-output GP Simulations

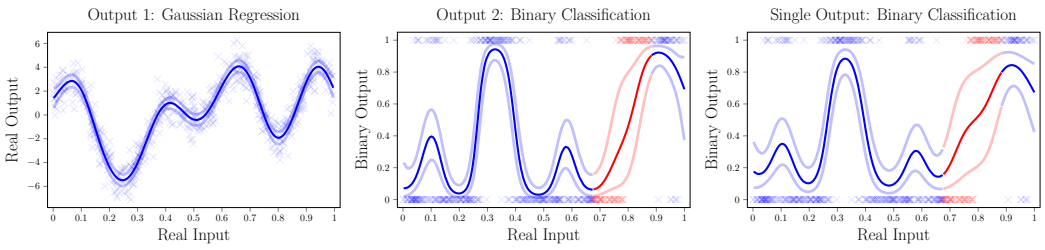
In this section, we evaluate the heterogeneous MOGP model on different scenarios with irregular data. To demonstrate its performance in terms of multi-output learning, prediction and scalability, we explored several applications with both synthetic and real-world data. For all the experimental results in this section, we considered the *vanilla* RBF kernel for each covariance function $k_q(\cdot, \cdot)$. The number of latent functions $u_q(\cdot)$ was set to $Q = 3$.

For the standard optimization, i.e. without the need of stochastic gradient updates, we used the LBFGS-B algorithm. When SVI was explicitly needed, we considered the ADADELTA method and mini-batch sizes of 500 samples per output. All the performance metrics are given in terms of the negative log-predictive density (NLPD), calculated from the test subsets of data. This sort of metrics are also applicable to any type of likelihood function and its computation is straightforward.

Missing Gap Prediction

In the first experiment, we evaluate if the model is able to predict observations in one output using training information from another one. We setup a toy problem which consists of $D = 2$ heterogeneous outputs, where the first function $y_1(\boldsymbol{x})$ is real-valued and $y_2(\boldsymbol{x})$ binary. Assumming that heterogeneous outputs do not share a common input set, we observe $N_1 = 600$ and $N_2 = 500$ samples respectively. All inputs are uniformly distributed in the input range $[0, 1]$, and we generate a *gap* only in the set of binary observations by removing $N_{\text{test}} = 150$ samples in the interval $[0.7, 0.9]$. Using the remaining points from both outputs for training, we fitted the heterogeneous MOGP model.

In Fig. 2.3, looking to subfigures (a) and (b), we see how the uncertainty in binary test predictions is reduced by learning from the first output, which is a regression problem. In contrast, Fig. 2.3 (c) shows wider variance in the predicted parameter when it is trained independently. For the multi-output case in the left-hand subfigures, we obtained an NLPD value on test data of $32.5 \pm 0.2 \times 10^{-2}$, while in the single-output case of the right-hand subfigure, the NLPD was $40.51 \pm 0.08 \times 10^{-2}$.



(a) (b) (c)

Figure 2.3: Comparison between multi-output and single-output performance for two heterogeneous sets of observations. (A) Fitted function and uncertainty for the first output. It represents the mean function parameter $\mu(\boldsymbol{x})$ for a Gaussian distribution with $\sigma^2 = 1$. (B) Predictive output function for binary inputs. Blue curve is the fitting function for training data and the red one corresponds to predicting from test inputs (true test binary outputs in red too). (C) Same output as in (B) but training an independent chained GP only in the single binary output (GP binary classification).

Human Behavior Learning

In this experiment, and also motivated by the final objective of this thesis, we are interested in modeling human behavior. Particularly, with application to psychiatric patients. Previous work by Soleimani et al. (2018) already explored the application of scalable MOGP models to healthcare for reliable predictions from multivariate temporal data. The data of this experiment comes from a medical study that asked patients to download a monitoring *app* (EB2) on their personal smartphones (Berrouiguet et al., 2018). The ubiquitous system is able to capture traces about mobility, communication metadata and interactions in social media networks. The work has a particular interest in mental health as we well see later in Ch. 5. Particularly, shifs or misalignment in the circadian feature (24h cycles) of the human behavioral patterns captured, can be interpreted as early signs of crisis.

	BERNOULLI	HETEROSCEDASTIC	BERNOULLI	GLOBAL
HETMOGP	$\underline{2.24 \pm 0.21}$	$\underline{6.09 \pm 0.21}$	5.41 ± 0.05	$\underline{13.74 \pm 0.41}$
CHGP	2.43 ± 0.30	7.29 ± 0.12	$\underline{5.19 \pm 0.81}$	14.91 ± 1.05

Table 2.1: Behavior dataset test-NLPD ($\times 10^{-2}$). Best NLPD metrics are underlined.

The preprocessing stage of this experiment consisted of three steps. First, we obtained a binary indicator variable of presence/absence at home by monitoring *latitude-longitude* and measuring its distance from the original patient's home location within a 50m radius range. All locations were previously anonymized in a privacy-preserving manner. Then, using the already measured distances, we generated a mobility sequence with all log-distance values. The last step was to obtain binary samples representing the use/non-use of the WHATSAPP application in the personal smartphone. At each monitoring time instant, we used its differential data consumption to determine the use/non-use of the application. Finally, we considered an entire week in seconds as the input domain $\boldsymbol{x}_n \in \mathcal{X}$, that we also normalized to the range $[0, 1]$.

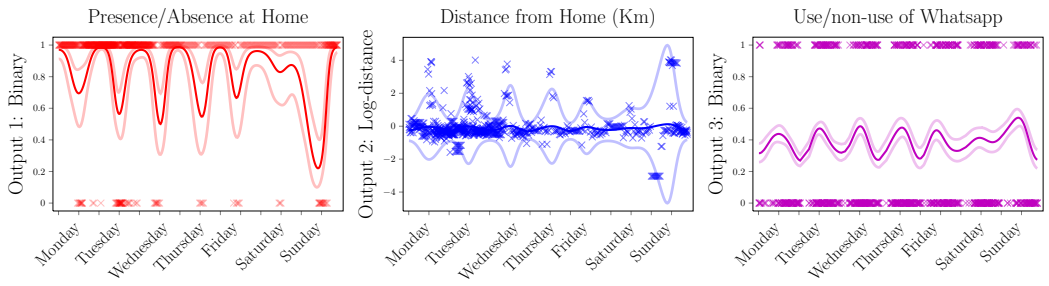


Figure 2.4: Results for multi-output modeling of human behavioral data. After the training process, all output predictions share a common (daily) periodic pattern, initially induced by the social media features (*purple*).

Demographic Modelling

Based on the large scale experiments in Hensman et al. (2013a), we obtained the complete register of housing properties sold in the Greater London County during 2017 (https:www.gov.uk/government/collections/price-paid-data). The data was preprocessed and particularly, we translated all property addresses to *latitude-longitude* coordinates. For each spatial input data-point, we considered two observations, one binary and one real-valued. The first one indicates if the property is or not a flat (zero would mean detached, semi-detached, terraced, etc..), and the second output is the sale price of houses. Particularly, we transformed this output to the real domain using the mapping $\log(y+1)$.

	Bernoulli	Heteroscedastic	Global
hetmogp	$\underline{6.38 \pm 0.46}$	$\underline{10.05 \pm 0.64}$	$\underline{16.44 \pm 0.01}$
chgp	6.75 ± 0.25	10.56 ± 1.03	17.31 ± 1.06

Table 2.2: Demographic modelling test-NLPD ($\times 10^{-2}$). Best NLPD metrics are underlined.

Our goal is to predict the heterogeneous features of houses given a certain location in the London area. We used a training set of $N = 20,000$ samples, and $1,000$ for test prediction. The number of inducing points was set to $M = 100$. Results in Fig. 2.5 show a portion of the entire heterogeneous dataset and its test prediction curves. We obtained a global NLPD score of $\underline{16.44 \pm 0.01}$ using the MOGP model and 17.31 ± 1.06 in the independent output setting; both metrics are ($\times 10^{-2}$). There is also an improvement in performance when training our multi-output model even in large scale datasets. Error scores are shown in Tab. 2.2 per output.

High-dimensional Inputs

In the last experiment with the heterogeneous MOGP model, we tested its robustness on the *arrythmia* dataset, available in the UCI repository (http://archive.ics.uci.edu/ml). We use a dataset of dimensionality $p = 255$ and 452 samples that we split in training, validation and test subsets. We use the model for predicting the binary output (gender) and the continuous output (logarithmic age). The prediction of both output values is compared against independent chained GPs, one per output function. The binary samples are modelled using a Bernoulli likelihood distribution and the continuous one as Gaussian distributed. We

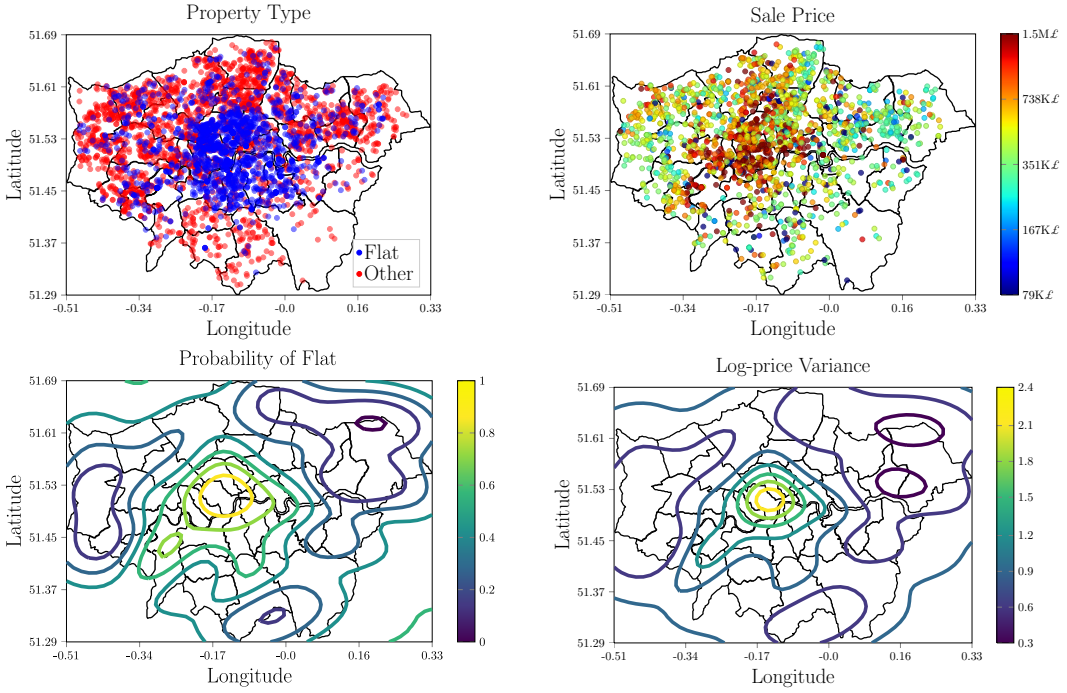


Figure 2.5: Results for spatial modeling of heterogeneous data with MOGP models. (TOP ROW) 10% of training samples for the *Greater London County*. Binary outputs in red-blue colors are the type of property sold in 2017 and real-valued ones are the log-prices included in the house sale contracts. (BOTTOM ROW) Posterior predictive test curves for $N_{\text{test}} = 2,500$ input samples. Multimodal $\mathbb{R}^2$ curves share common patterns as long as their LPFs are linear combinations of a unique set of components $\{u_q\}_{q=1}^{Q}$.

obtained an average NLPD value of 0.0191 for both the multi-output and independent output function models. The slight difference was in the standard deviation, which was better with the presented methodology.

2.3.3 Heterogeneous Single-output GP Simulations

Based on Moreno-Muñoz et al. (2020a) and the local GP models that will be later introduced in the next Ch. 3, we analysed how single-output heterogeneous GP priors can be used if one of the tasks is of *regression* type (real-valued continuous outputs) and the other a GP *classification* (binary samples) problem.

For the simulations depicted in Fig. 2.6, we considered the demographic dataset also used in Sec. 2.3.2 for spatial modelling with MOGPs. In this new case, we have two subsets of input-output pairs of variables: i) the binary contract of houses (*leasehold* vs. *freehold*) and ii) the log-price per latitude-longitude coordinate in the Greater London County. Looking to Fig. 2.6, we show the input space divided into four quadrants (Q) of the city area{Q1, Q2, Q3, Q4}. Data points contained in the quadrant Q1 are exclusively real-valued and Gaussian distributed. On the other hand, quadrants {Q2, Q3, Q4} are trained with a different GP classifier, and output variables are binary. To clarify, Q1 is the right-upper corner given the central axes. Particularly, our purpose in this experiment is to learn a single latent GP function from the binary data on {Q2, Q3, Q4} via classification and Q1 via regression.

This is, the same function f parameterizes different likelihood densities at each subset of quadrants.

To evaluate the predictive performance of the proposed inference model, we *hid* the latent process f in order to be predicted within a Bernoulli likelihood density in Q1. In particular, we want to asses whether the GP function f, learned partially with real-valued data from house prices on the other quadrants, can improve the prediction task (classification) over the local area where we first learned only with binary observations from Q1. The heterogeneous model shows a test NLPD of 7.94 ± 0.01 in independent classification while the joint task predicts with an NLPD of 8.00 ± 0.01 in the Q1. We asses that the heterogeneous GP prediction is better in Q1 than the independent local GP classifier. This result shows a similar advance as the ones obtained with the MOGP model in Sec. 2.3.2, shown in Fig. 2.5. The mean function of the GP regressor is passed through the *sigmoid* function to clarify the multimodality in Fig. 2.6.

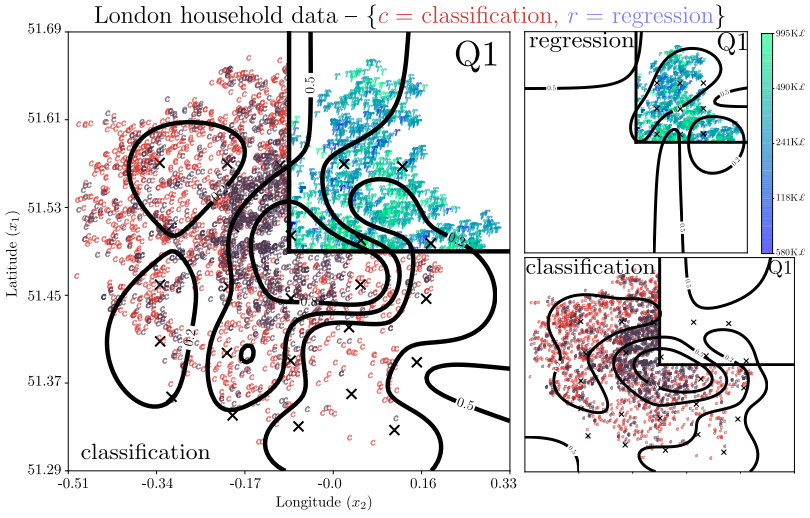


Figure 2.6: Results for the spatial GP modeling of a single-output heterogeneous data collection. (LEFT COLUMN) 75% of training samples are binary and the rest are real-valued and Gaussian distributed. The demographic dataset corresponds to the *Greater London County*. Binary outputs (red) are the type of property sold in 2017 and real ones (green-blue) are prices included in sale contracts. (RIGHT COLUMN) Test prediction curves for the local regression and classification tasks. Prices of houses are in log-scale.

2.4 Discussion

This chapter introduces two novel extensions of well-known statistical methods for handling heterogenous data collections. Based on the standard latent class models, whose discrete latent structure is of particular interest later in Ch. 4 and 5, we developed an heterogeneous circadian mixture. This model can be used to capture the uncertainty from arbitrary combinations of statistical data types and, at the same time, characterize the underlying short-term periodicities between the features of observation vectors. Additionally, we introduced covariance functions to accept non-stationary periodic samples, whose components are restricted to 24h, but also accept others. Despite that the model introduces conditioning

across heterogenous likelihood densities via the discrete latent assignments, the parameterization of attributes is still independent. In contrast to this assumption of independence, we extended multi-output GP models for heterogeneous observations. The GP model is able to work on large scale datasets by using sparse approximations within stochastic variational inference methods. In this case, we illustrated how a linear model of corregionalisation (LMC) correlates all the output parameter functions in the densities. Experimental results show promising improvements with respect to the *independent* learning of heterogeneous data attributes in both settings. One based on latent variables and another on non-linear parameterizations of densities with stochastic processes.

CHAPTER 3

CONTINUAL AND DISTRIBUTED INFERENCE

THE first issue that a probabilistic model must face when addressing the problem of human behavior learning is the observation of heterogeneous statistical variables. However, having addressed this particular problem in the preceding Ch. 2, we now focus on the manner that the irregular data are observed, its influence in the inference process of models or the methods that we use for fitting them. In the particular case of this chapter, we pay attention to the way of accessing observations, going one step further than the traditional assumption. This one is typically related to the case where all samples can be *revisited* without restrictions *a priori*. In practice, when modelling human behavior, one encounter examples where the massive storage or data centralisation is not possible anymore for preserving the privacy of individuals, e.g. healthcare or behavioral data. Another potential restriction would be the need of providing an output prediction rather than awaiting for the entire observation of the desired event. For instance, it would be meaningless to await three or five years until sufficient data from a mental health patient would be stored. Sometimes, the decision-making steps based on the probabilistic predictions cannot be delayed. The mere limitation of the data availability forces learning algorithms to derive novel capabilities, such as i) distributing the data for *federated learning*, ii) observe streaming samples for *continual learning* and iii) limiting data exchange for *private-owned models*.

A common theme in the previous approaches is the idea of model memorising and recycling, i.e. using the already fitted parameters in another problem or joining them to others for an additional global task without revisiting any data. If we look to the functional view of this idea, uncertainty metrics are still much harder to be repurposed than parameters. This is the point where stochastic processes, and particularly Gaussian process models, once again, play their role in this thesis.

We begin in Sec. 3.1 with the presentation of *recyclable* Gaussian process models, a novel methodology for distributing the computational cost of inference accross several nodes or machines. The main point of interest in this model is the ability for re-building new global tasks from the subsets of distributed parameters without revisiting any data. The framework ensembles independent variational approximations of Gaussian processes and allows for regression, classification or even single-output heterogeneous likelihoods, as in the previous Sec. 2.2.3. A similar spirit is preserved in Sec. 3.2, where we address the problem of continual learning in both single-output and multi-task Gaussian process models. The general idea is to extend Gaussian processes for handling sequential input-output observations. The approach uses the existing prior-posterior recursion of online Bayesian inference, i.e. where past posterior discoveries become future prior beliefs, to the infinite-dimensional view of function spaces.

The technical results described in this chapter have been previously presented in two main pieces of work. The first one, Moreno-Muñoz et al. (2020a) was submitted to the 2020 Conference on Advances in Neural Information Processing Systems (NeurIPS), and now it is been revised for a future re-submission. The second piece of work, Moreno-Muñoz et al. (2019) was initially submitted to the Journal of Machine Learning Research (JMLR) and a revised version of the manuscript will be submitted in the near future.

3.1 Recyclable Gaussian Processes

In this section, we investigate the novel approach of a general framework for recycling distributed variational *sparse* approximations to Gaussian process (GP) models, as illustrated in the Fig. 3.1. It based on the properties of the Kullback-Leibler divergence between stochastic processes (Matthews et al., 2016) and the formal definition of Bayesian inference. The proposed *recyclable* method ensembles an arbitrary amount of variational GP models with different complexity, likelihood and location of pseudo-inputs, without revisiting any data.

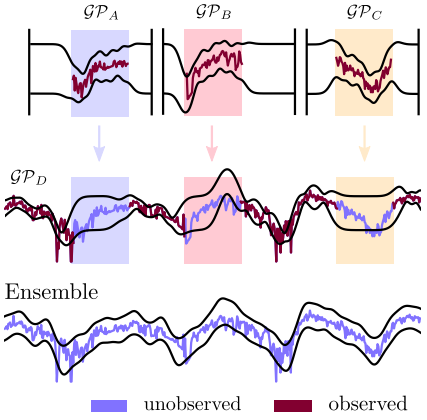


Figure 3.1: Recyclable GPs (A, B, C and D) are re-combined without accessing to the subsets of observations.

Data Formulation

Similarly as we did in the preceding Ch. 2, we consider a supervised learning problem where we observe input-output training data $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^N$, with $\boldsymbol{x} \in \mathbb{R}^p$. The general assumption is to consider that output samples y_n are i.i.d. and can be either discrete or continuous variables. In this section, we also express the likelihood model as $p(y|f)$, being the non-linear function $f(\cdot)$ the one that is generated from a zero-mean GP prior of the form $f \sim \mathcal{GP}(0, k(\cdot,\cdot))$. The covariance function $k(\cdot,\cdot)$ is chosen as the *vanilla* kernel (RBF) in this section, but many others can be also introduced without any restriction.

The dataset $\mathcal{D}$ is assumed to be partitioned into an arbitrary number (i.e. 2, 5, 10, 100, 1K, etc) of K subsets that we aim to observe and process independently, that is, $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K\}$. There are no restrictions on the amount of subsets or the number of learning nodes to be used. The subsets $\{\mathcal{D}_k\}_{k=1}^K$ do not need to be of the same size, and we only restrict them to be $N_k < N$. However, since we think in applications with a large number of observations, even per subset $\mathcal{D}_k$, e.g. of a patient, we still consider that N_k for all $\{1, 2, \ldots, K\}$ is sufficiently large for not accepting exact GP inference due the temporal and computational demand of the inversion of covariance matrices. Notice that the variable k is an index now, while $k(\cdot,\cdot)$ makes reference to the *kernel* function of the GP model.

3.1.1 Sparse Approximations for Distributed Subsets

In this recyclable approach, we also adopt the sparse GP methodology based on the *inducing variables*, together with the variational framework of Titsias (2009b). In our particular formulation with K distributed partitions and their adjacent samples, we define subsets of $M_k \ll N_k$ inducing inputs, such that $\boldsymbol{Z}_k = \{\boldsymbol{z}_m\}_{m=1}^{M_k}$, where $\boldsymbol{z}_m \in \mathbb{R}^p$ and their non-linear function evaluations by $f(\cdot)$ are denoted as $\boldsymbol{u}_k = [f(\boldsymbol{z}_1), f(\boldsymbol{z}_2), \cdots, f(\boldsymbol{z}_{M_k})]^\top$. We remark that f is considered to be stationary across all distributed tasks, being $\boldsymbol{u}_k$ $\forall k \in [1, 2, \ldots, K]$ values of the same function. A detailed review of this type of variational sparse approximations is included above, in Ch. 2.

Here, to obtain multiple independent approximations to the posterior distribution $p(f|\mathcal{D})$ of the GP function, we introduce K auxiliary variational densities $q_k(f)$, one per distributed partition $\mathcal{D}_k$. Additionally, each variational $q_k(f)$ on every k-th local domain factorises according to

$$q_k(f) = p(f_{\neq \boldsymbol{u}_k}|\boldsymbol{u}_k)q_k(\boldsymbol{u}_k), \tag{3.1}$$

with $q_k(\boldsymbol{u}_k) = \mathcal{N}(\boldsymbol{u}_k|\boldsymbol{\mu}_k, \boldsymbol{S}_k)$ and $p(f_{\neq \boldsymbol{u}_k}|\boldsymbol{u}_k)$ being the standard conditional GP prior distribution given the set of hyperparameters $\boldsymbol{\gamma}_k$ for each kth *kernel* function. Importantly, to fit the local variational densities $q_k(\boldsymbol{u}_k)$, we build lower bounds $\mathcal{L}_k$ on the marginal log-likelihood distribution (ELBO) of every data partition $\mathcal{D}_k$. Then, we use optimisation methods, typically gradient-based, to maximise the K objective functions $\mathcal{L}_k$ in an *asynchronous* manner. One per distributed task, separately. Each *local* ELBO is obtained as follows

$$\mathcal{L}_k = \sum_{n=1}^{N_k} \mathbb{E}_{q_k(\boldsymbol{f}_n)} \left[\log p(y_n|\boldsymbol{f}_n)\right] - \mathrm{KL}[q_k(\boldsymbol{u}_k)||p_k(\boldsymbol{u}_k)], \tag{3.2}$$

with $p_k(\boldsymbol{u}_k) = \mathcal{N}(\boldsymbol{u}_k|\boldsymbol{0}, \mathbf{K}_k)$, where covariance matrices $\mathbf{K}_k \in \mathbb{R}^{M_k \times M_k}$ have entries $k(\boldsymbol{z}_m, \boldsymbol{z}'_m)$ with $\boldsymbol{z}_m, \boldsymbol{z}_{m'} \in \boldsymbol{Z}_k$. They are also conditioned to certain kernel hyperparameters $\boldsymbol{\gamma}_k$ that we also aim to estimate. The *bold* variable $\boldsymbol{f}_n$ corresponds to instances $f(\boldsymbol{x}_n)$ and the marginal posterior density comes from the following integration

$$q_k(\boldsymbol{f}_n) = \int p(\boldsymbol{f}_n|\boldsymbol{u}_k) q_k(\boldsymbol{u}_k) d\boldsymbol{u}_k. \tag{3.3}$$

In practice, the distributed local bounds $\mathcal{L}_k$ are identical to the one presented in Hensman et al. (2015a) and also accept stochastic variational inference (Hoffman et al., 2013; Hensman et al., 2013a). An important detail on this derivations is that, while the GP function is restricted to be *stationary* across tasks, the likelihood distribution model $p(y_n|\boldsymbol{f}_n)$ is not. This point opens the door to the heterogeneous likelihood setting presented in Sec. 2.2.3.

3.1.2 Global Inference from Local Learning

Having a *dictionary* which contains the already fitted local variational solutions, while others can be still under computation, we focus on how using them for performing global inference of the GP model. Such dictionary consists, for instance, of a list of objects $\mathcal{E} = \{\mathcal{E}_1, \mathcal{E}_2, \cdots, \mathcal{E}_K\}$ without any specific order, where each object are the parameters $\mathcal{E}_k = \{\boldsymbol{\phi}_k, \boldsymbol{\gamma}_k, \boldsymbol{Z}_k\}$. We use $\boldsymbol{\phi}_k$ to denote the variational parameters, e.g. $\boldsymbol{\mu}_k$ and $\boldsymbol{S}_k$.

Ideally, to obtain a global inference solution given the GP models described in the dictionary, the resulting posterior distribution should be valid for all the local subsets of data. In practice, this is only possible if we consider the entire dataset $\mathcal{D}$ in a *maximum likelihood criterion* setting. Specifically, our goal now is to obtain an approximate posterior distribution $q(f) \approx p(f|\mathcal{D})$ by maximising a lower bound $\mathcal{L}_\mathcal{E}$ under the log-marginal likelihood $\log p(\mathcal{D})$ over *all* observations without revisiting the data already observed by the local models.

For this task, we begin by considering the *full* posterior distribution of the stochastic process, similarly as Burt et al. (2019) does for obtaining an upper bound on the KL divergence. The principal idea is to use infinite-dimensional integral operators, previously introduced by Matthews et al. (2016) in the context of variational inference, and even before by Seeger (2002) for standard GP error bounds.

The use of infinite-dimensional integrals is equivalent to an *augment-and-reduce* strategy (Ruiz et al., 2018). It consists of two main steps: i) we augment the model to accept the conditioning on the infinite-dimensional stochastic process $f(\cdot)$ and ii) we use properties of Gaussian marginals to reduce the infinite-dimensional integral operators to a finite amount of GP function values of interest. Similar strategies have been used in the context of continual learning for GPs (Bui et al., 2017a; Moreno-Muñoz et al., 2019). This will be later discussed in Sec. 3.2.

Global Objective Function

The variational construction considered is as follows. We first denote $\boldsymbol{y}$ as all the output targets $\{y_n\}_{n=1}^N$ in the dataset $\mathcal{D}$ and f_∞ as the *augmented* infinite-dimensional GP function. Notice that we assume that f_∞ contains all the function values taken by $f(\cdot)$, including that instances at $\{\boldsymbol{x}_n\}_{n=1}^{N_k}$ and $\{\boldsymbol{Z}_k\}_{k=1}^K$ for all the k-th partitions.

The augmented log-marginal expression is therefore

$$\log p(\boldsymbol{y}) = \log p(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_K) = \log \int p(\boldsymbol{y}, f_\infty) df_\infty, \tag{3.4}$$

where each $\boldsymbol{y}_k = \{y_n\}_{n=1}^{N_k}$ is the subset of output values already used for training the local GP modes in Eq. (3.2). The joint distribution expanded via integration in Eq. (3.4) factorises according to

$$\log \int p(\boldsymbol{y}, f_\infty) df_\infty = \log \int p(\boldsymbol{y}|f_\infty) p(f_\infty) df_\infty, \tag{3.5}$$

where $p(\boldsymbol{y}|f_\infty)$ is the *augmented* likelihood term of all output targets of interest and $p(f_\infty)$ the GP prior over the infinite amount of points in the input-space $\mathbb{R}^p$. This last density takes the form of an infinite-dimensional Gaussian, that we will avoid to evaluate explicitly in the following equations.

To build the lower bound on the log-marginal distribution shown in Eq. (3.4), we introduce a *global* variational distribution $q(\boldsymbol{u}_*) = \mathcal{N}(\boldsymbol{u}_*|\boldsymbol{\mu}_*, \boldsymbol{S}_*)$ that we aim to fit. The inducing variables $\boldsymbol{u}_*$ correspond to function instances of the process $f(\cdot)$ given the new subset of inducing inputs $\boldsymbol{Z}_* = \{\boldsymbol{z}_m\}_{m=1}^M$, where M is the free-complexity degree of this global variational distribution. The expression of the log-marginal distribution within the new density $q(\boldsymbol{u}_*)$ is

$$\begin{aligned}\log p(\boldsymbol{y}) = \log \int p(\boldsymbol{y}|f_\infty) p(f_\infty) df_\infty &= \log \int \frac{q(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} p(\boldsymbol{y}|f_\infty) p(f_\infty) df_\infty \\ &= \log \iint \frac{q(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} p(\boldsymbol{y}|f_\infty) p(f_{\infty \neq \boldsymbol{u}_*}|\boldsymbol{u}_*) p(\boldsymbol{u}_*) df_{\infty \neq \boldsymbol{u}_*} d\boldsymbol{u}_*. \end{aligned} \tag{3.6}$$

Notice that the differentials df_∞ have been splitted into $df_{\infty \neq \boldsymbol{u}_*} d\boldsymbol{u}_*$, and at the same time, we applied properties of Gaussian conditionals in the GP prior to re-write the distribution $p(f_\infty)$ as the product $p(f_{\infty \neq \boldsymbol{u}_*}|\boldsymbol{u}_*)p(\boldsymbol{u}_*)$.

To derive the preliminary bound in Eq. (3.6), we also exploited the reparameterisation trick introduced by Gal et al. (2014) for distributing the computational load in the main expectation term of variational inference methods for GPs. It is based on a double-application of the Jensen's inequality, one w.r.t. the new instances $\boldsymbol{u}_*$ and another w.r.t. the rest of function values $f_{\infty \neq \boldsymbol{u}_*}$. It is obtained as

$$\begin{aligned}\log p(\boldsymbol{y}) &= \log \iint \frac{q(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} p(\boldsymbol{y}|f_\infty) p(f_{\infty \neq \boldsymbol{u}_*}|\boldsymbol{u}_*) p(\boldsymbol{u}_*) df_{\infty \neq \boldsymbol{u}_*} d\boldsymbol{u}_* \\ &= \log \iint q(\boldsymbol{u}_*) p(f_{\infty \neq \boldsymbol{u}_*}|\boldsymbol{u}_*) p(\boldsymbol{y}|f_\infty) \frac{p(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} df_{\infty \neq \boldsymbol{u}_*} d\boldsymbol{u}_* \\ &= \log \left(\mathbb{E}_{q(\boldsymbol{u}_*)} \left[\mathbb{E}_{p(\boldsymbol{f}_{\infty \neq \boldsymbol{u}_*}|\boldsymbol{u}_*)} \left[p(\boldsymbol{y}|f_\infty) \frac{p(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} \right] \right] \right) \\ &\geq \mathbb{E}_{q(\boldsymbol{u}_*)} \left[\log \left(\mathbb{E}_{p(\boldsymbol{f}_{\infty \neq \boldsymbol{u}_*}|\boldsymbol{u}_*)} \left[p(\boldsymbol{y}|f_\infty) \frac{p(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} \right] \right) \right] \\ &\geq \mathbb{E}_{q(\boldsymbol{u}_*)} \left[\mathbb{E}_{p(\boldsymbol{f}_{\infty \neq \boldsymbol{u}_*}|\boldsymbol{u}_*)} \left[\log \left(p(\boldsymbol{y}|f_\infty) \frac{p(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} \right) \right] \right]. \end{aligned} \tag{3.7}$$

Moreover, the last prior distribution is $p(\boldsymbol{u}_*) = \mathcal{N}(\boldsymbol{u}_*|\mathbf{0}, \mathbf{K}_{**})$ where $[\mathbf{K}_{**}]_{m,n} := k(\boldsymbol{z}_m, \boldsymbol{z}_n)$, with $\boldsymbol{z}_m, \boldsymbol{z}_n \in \boldsymbol{Z}_*$, conditioned to the *global* kernel hyperparameters $\boldsymbol{\psi}_*$ that we also want to estimate. Notice that the double expectation integration is placed in the last two terms of Eq. (3.6), and it comes from the factorization of the infinite-dimensional integral operator and the application of the Jensen's inequality twice. The full derivation of this bound can be found in the appendix of this thesis.

The *nested* integration in the bound in Eq. (3.7) allows to ensemble the local GP tasks via the likelihood densities. This is the goal of the following sections.

Local Likelihood Reconstruction

The *augmented* likelihood distribution $p(\boldsymbol{y}|f_\infty)$ is, perhaps, the most important ingredient of our derivations in this section. It allows us to apply conditional independence (CI) between the subsets of distributed output targets $\boldsymbol{y}_k$. This gives a factorized term that we later use for introducing the local variational experts in the bound in Eq. (3.7), that is

$$\log p(\boldsymbol{y}|f_\infty) = \sum_{k=1}^{K} \log p(\boldsymbol{y}_k|f_\infty). \tag{3.8}$$

To avoid the revisiting of *old* local likelihood terms, and hence, re-evaluating the distributed subsets of data $\mathcal{D}_k$ that might not be available anymore, we use the Bayes theorem but conditioned to the infinite-dimensional augmentation in Eq. (3.8). It indicates that the local variational densities $q_k(f_\infty)$ can be approximated as

$$q_k(f_\infty) \approx p(f_\infty|\boldsymbol{y}_k) \propto p(f_\infty)p(\boldsymbol{y}_k|f_\infty), \tag{3.9}$$

where the augmented approximate distribution also factorises according to

$$q_k(f_\infty) = p(f_{\infty \neq \boldsymbol{u}_k}|\boldsymbol{u}_k)q_k(\boldsymbol{u}_k), \tag{3.10}$$

as in the variational framework used in Titsias (2009b). Similar expressions consisting on the full stochastic process conditionals were previously used in Bui et al. (2017a) and Matthews et al. (2016), with a particular emphasis on the theoretical consistency of the augmentation.

Thus, we can find a reliable approximation for each local likelihood term $p(\boldsymbol{y}_k|f_\infty)$ by inverting the Bayes theorem in Eq. (3.9). Then, the double conditional expectation in Eq. (3.7) turns to be

$$\begin{aligned}\mathbb{E}_{p(\boldsymbol{f}_{\infty \neq \boldsymbol{u}_*}|\boldsymbol{u}_*)}\left[\log p(\boldsymbol{y}|f_\infty)\right] &\approx \sum_{k=1}^{K} \mathbb{E}_{p(\boldsymbol{f}_{\infty \neq \boldsymbol{u}_*}|\boldsymbol{u}_*)}\left[\log \frac{q_k(f_\infty)}{p(f_\infty)}\right] \\ &= \sum_{k=1}^{K} \mathbb{E}_{p(\boldsymbol{u}_k|\boldsymbol{u}_*)}\left[\log \frac{q_k(\boldsymbol{u}_k)}{p(\boldsymbol{u}_k)}\right], \end{aligned} \tag{3.11}$$

where we applied properties of Gaussian marginals to *reduce* the infinite-dimensional expectation, and factorised the process densities to be explicit on each subset of fixed inducing-variables $\boldsymbol{u}_k$ rather than all f_∞ instances. For example, the integral $\int p(f_\infty) df_{\infty \neq \boldsymbol{u}_k}$ is analogous to $\int p(\boldsymbol{f}_{\infty \neq \boldsymbol{u}_k}, \boldsymbol{u}_k) d\boldsymbol{f}_{\infty \neq \boldsymbol{u}_k} = p(\boldsymbol{u}_k)$ via marginalisation.

Variational Contrastive Expectations

The introduction of K expectation terms over the log-ratios in the bound of Eq. (3.7), as a substitution of the local likelihoods, leads to particular advantages in the model. If we have a

nested integration in Eq. (3.7), first over $\boldsymbol{u}_*$ at the conditional prior distribution, and second, over $\boldsymbol{u}_k$ given the log-ratio $\log(q_k(\boldsymbol{u}_k)/p(\boldsymbol{u}_k))$, we can exploit the GP predictive equation to write down

$$\sum_{k=1}^{K} \mathbb{E}_{q(\boldsymbol{u}_*)} \left[\mathbb{E}_{p(\boldsymbol{u}_k|\boldsymbol{u}_*)} \left[\log \frac{q_k(\boldsymbol{u}_k)}{p(\boldsymbol{u}_k)} \right] \right] = \sum_{k=1}^{K} \mathbb{E}_{q_{\mathcal{C}}(\boldsymbol{u}_k)} \left[\log \frac{q_k(\boldsymbol{u}_k)}{p(\boldsymbol{u}_k)} \right], \tag{3.12}$$

where we first obtained $q_{\mathcal{C}}(\boldsymbol{u}_k)$ via the integral

$$q_{\mathcal{C}}(\boldsymbol{u}_k) = \int q(\boldsymbol{u}_*) p(\boldsymbol{u}_k|\boldsymbol{u}_*) d\boldsymbol{u}_*, \tag{3.13}$$

that coincides with approximate predictive GP posterior. This distribution can be obtained analytically for each kth subset $\boldsymbol{u}_k$ using the following expression, whose complete derivation is provided in the appendix of this thesis. That is,

$$q_{\mathcal{C}}(\boldsymbol{u}_k) = \mathcal{N}(\boldsymbol{u}_k|\mathbf{K}_{*k}^{\top}\mathbf{K}_{**}^{-1}\boldsymbol{\mu}_*, \mathbf{K}_{kk} + \mathbf{K}_{*k}^{\top}\mathbf{K}_{**}^{-1}(\boldsymbol{S}_* - \mathbf{K}_{**})\mathbf{K}_{**}^{-1}\mathbf{K}_{*k}), \tag{3.14}$$

where, once again, we use $\boldsymbol{\phi}_* = \{\boldsymbol{\mu}_*, \boldsymbol{S}_*\}$, the global variational parameters that we aim to learn. One important detail of the sum of expectations in Eq. (3.12) is that it works as an average contrastive indicator that measures how well the global distribution $q(\boldsymbol{u}_*)$ is being fitted to the local experts $q_k(\boldsymbol{u}_k)$.

Without the need of revisiting any of the previously observed subsets of data, with their adjacent distributed samples, the GP predictive distribution $q_{\mathcal{C}}(\boldsymbol{u}_k)$ is playing a different role in contrast to the usual one. Typically, we assume the approximate posterior $q(\cdot)$ fixed and fitted, and we evaluate its predictive performance on some test data points. In this case, it goes in the opposite way. The approximate variational distribution $q_{\mathcal{C}}(\boldsymbol{u}_k)$ is unfixed, and it is instead evaluated over each k-th local subset of inducing-inputs $\boldsymbol{Z}_k$.

Lower Ensemble Bounds

We are now able to simplify the initial bound in Eq. (3.7) by direct substitution of the first term with the contrastive expectations presented in Eq. (3.12). This substitution gives us the final version of the lower bound $\mathcal{L}_{\mathcal{E}} \leq \log p(\boldsymbol{y})$ on the log-marginal likelihood for the global GP, that is

$$\begin{aligned}
\mathcal{L}_{\mathcal{E}} &= \mathbb{E}_{q(\boldsymbol{u}_*)} \left[\sum_{k=1}^{K} \mathbb{E}_{p(\boldsymbol{u}_k|\boldsymbol{u}_*)} \left[\log \left(\frac{q_k(\boldsymbol{u}_k)}{p_k(\boldsymbol{u}_k)} \right) \right] - \log \left(\frac{q(\boldsymbol{u}_*)}{p(\boldsymbol{u}_*)} \right) \right] \\
&= \sum_{k=1}^{K} \mathbb{E}_{q(\boldsymbol{u}_*)} \left[\mathbb{E}_{p(\boldsymbol{u}_k|\boldsymbol{u}_*)} \left[\log \left(\frac{q_k(\boldsymbol{u}_k)}{p_k(\boldsymbol{u}_k)} \right) \right] \right] - \mathbb{E}_{q(\boldsymbol{u}_*)} \left[\log \left(\frac{q(\boldsymbol{u}_*)}{p(\boldsymbol{u}_*)} \right) \right] \\
&= \sum_{k=1}^{K} \mathbb{E}_{q(\boldsymbol{u}_*)} \left[\mathbb{E}_{p(\boldsymbol{u}_k|\boldsymbol{u}_*)} \left[\log \left(\frac{q_k(\boldsymbol{u}_k)}{p_k(\boldsymbol{u}_k)} \right) \right] \right] - \mathrm{KL}\left[q(\boldsymbol{u}_*) || p(\boldsymbol{u}_*) \right] \\
&= \sum_{k=1}^{K} \mathbb{E}_{q_{\mathcal{C}}(\boldsymbol{u}_k)} \left[\log q_k(\boldsymbol{u}_k) - \log p_k(\boldsymbol{u}_k) \right] - \mathrm{KL}\left[q(\boldsymbol{u}_*) || p(\boldsymbol{u}_*) \right],
\end{aligned} \tag{3.15}$$

The maximisation of this bound $\mathcal{L}_{\mathcal{E}}$ is w.r.t. the parameters $\boldsymbol{\phi}_*$, the hyperparameters $\boldsymbol{\psi}_*$ and the inducing inputs $\boldsymbol{Z}_*$. To assure the positive-definiteness of the variational covariance matrices $\{\boldsymbol{S}_k\}_{k=1}^{K}$ and $\boldsymbol{S}_*$ on both local and global cases, we consider that they all factorize

according to the Cholesky decomposition $\boldsymbol{S}_* = \mathbf{L}\mathbf{L}^\top$. We can then introduce unconstrained optimization methods to final the optimal values for the lower-triangular matrices $\mathbf{L}$.

A priori, the ensemble GP bound is *agnostic* with respect to the likelihood model chosen. There is also a general derivation in Matthews et al. (2016) about how stochastic processes and their integral operators are affected by projection functions. That is, different linking mappings of the function $f(\cdot)$ to the parametric domain of $\boldsymbol{\theta}$. In such cases, the local lower bounds $\mathcal{L}_k$ in Eq. (3.2) might include expectation terms that are just intractable. Since we build the *recyclable* framework to accept any possible statistical data-type, we propose to solve the integrals via Gaussian-Hermite quadratures are was originally done in Hensman et al. (2015a); Saul et al. (2016) and if it is not possible, an alternative would be to apply Monte-Carlo (MC) methods. As a reference, we included a little section in Ch. 2 on how quadratures can be used to solve expectation integrals in the context of variational GP models.

Gaussian Marginals for Infinite-dimensional Integral Operators

The generalized use of the properties of Gaussian marginals is the key point in the present section. In particular, such properties indicate that, having two *normal*-distributed random variables $\boldsymbol{a}$ and $\boldsymbol{b}$, its joint probability distribution function (pdf) is given by

$$p(\boldsymbol{a}, \boldsymbol{b}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb}\end{bmatrix}\right), \tag{3.16}$$

and if we want to marginalize out one of the variables $\{\boldsymbol{a}, \boldsymbol{b}\}$, then it turns to be

$$\int p(\boldsymbol{a}, \boldsymbol{b}) d\boldsymbol{b} = p(\boldsymbol{a}) = \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}). \tag{3.17}$$

This same property is applicable to every derivation with GP models and, in our case, it is the one that we use to reduce the infinite-dimensional integral operators over the full stochastic processes. An example of this can be found in the expectation terms $\mathbb{E}_{p(\boldsymbol{f}_{\infty \neq \boldsymbol{u}_*}|\boldsymbol{u}_*)}[\cdot]$ that we previously presented in Eq. (3.7). Its final derivation included in Eq. (3.15) to only integrate on $\boldsymbol{u}_k$ rather than on the function values $\boldsymbol{f}_{\infty \neq \boldsymbol{u}_*}$ comes from

$$\begin{aligned} p(\boldsymbol{f}_{\infty \neq \boldsymbol{u}_*}|\boldsymbol{u}_*) &= p(\boldsymbol{f}_{\infty \neq \{\boldsymbol{u}_*, \boldsymbol{u}_k\}}, \boldsymbol{u}_k|\boldsymbol{u}_*) \\ &= \mathcal{N}\left(\begin{bmatrix}\boldsymbol{m}_{\boldsymbol{f}_{\infty \neq \{\boldsymbol{u}_*, \boldsymbol{u}_k\}}|\boldsymbol{u}_*} \\ \boldsymbol{m}_{\boldsymbol{u}_k|\boldsymbol{u}_*}\end{bmatrix}, \begin{bmatrix}\boldsymbol{Q}_{\boldsymbol{f}_{\infty \neq \{\boldsymbol{u}_*, \boldsymbol{u}_k\}}|\boldsymbol{u}_*} & \boldsymbol{Q}_{\boldsymbol{f}_{\infty \neq \{\boldsymbol{u}_*, \boldsymbol{u}_k\}}, \boldsymbol{u}_k|\boldsymbol{u}_*} \\ \boldsymbol{Q}_{\boldsymbol{u}_k, \boldsymbol{f}_{\infty \neq \{\boldsymbol{u}_*, \boldsymbol{u}_k\}}|\boldsymbol{u}_*} & \boldsymbol{Q}_{\boldsymbol{u}_k|\boldsymbol{u}_*}\end{bmatrix}\right), \end{aligned} \tag{3.18}$$

that if we marginalize if we marginalize over $\boldsymbol{f}_{\infty \neq \{\boldsymbol{u}_*, \boldsymbol{u}_k\}}|\boldsymbol{u}_*$, ends in the following reduction of the conditional prior expectation

$$\begin{aligned} \mathbb{E}_{p(\boldsymbol{f}_{\infty \neq \boldsymbol{u}_*}|\boldsymbol{u}_*)}[g(\boldsymbol{u}_k)] &= \int p(\boldsymbol{f}_{\infty \neq \boldsymbol{u}_*}|\boldsymbol{u}_*) g(\boldsymbol{u}_k) d\boldsymbol{f}_{\infty \neq \boldsymbol{u}_*} \\ &= \iint p(\boldsymbol{f}_{\infty \neq \{\boldsymbol{u}_*, \boldsymbol{u}_k\}}, \boldsymbol{u}_k|\boldsymbol{u}_*) g(\boldsymbol{u}_k) d\boldsymbol{f}_{\infty \neq \{\boldsymbol{u}_*, \boldsymbol{u}_k\}} d\boldsymbol{u}_k \\ &= \int p(\boldsymbol{u}_k|\boldsymbol{u}_*) g(\boldsymbol{u}_k) d\boldsymbol{u}_k = \mathbb{E}_{p(\boldsymbol{u}_k|\boldsymbol{u}_*)}[g(\boldsymbol{u}_k)], \end{aligned} \tag{3.19}$$

where we also denoted $g(\boldsymbol{u}_k) = \log\left(q_k(\boldsymbol{u}_k)/p_k(\boldsymbol{u}_k)\right)$ and we also used

$$\int p(\boldsymbol{f}_{\infty \neq \{\boldsymbol{u}_*, \boldsymbol{u}_k\}}, \boldsymbol{u}_k|\boldsymbol{u}_*) d\boldsymbol{f}_{\infty \neq \{\boldsymbol{u}_*, \boldsymbol{u}_k\}} = p(\boldsymbol{u}_k) = \mathcal{N}(\boldsymbol{m}_{\boldsymbol{u}_k|\boldsymbol{u}_*}, \boldsymbol{Q}_{\boldsymbol{u}_k|\boldsymbol{u}_*}), \tag{3.20}$$

for the final derivation.

Computational Cost and Connections

The computational cost of the local models is $\mathcal{O}(N_k M_k^2)$, while the global GP reduces to $\mathcal{O}((\sum_k M_k)M^2)$ and $\mathcal{O}(M)$ in both training and prediction, respectively. The methods included in Tab. 3.1 typically need $\mathcal{O}(\sum_k N_k^2)$ for global prediction of tasks.

A last theoretical aspect is the link established between the global bound in Eq. (3.15) and the underlying idea introduced in Tresp (2000) and Deisenroth and Ng (2015). Distributed GP models are based on the application of CI to factorise the likelihood function across terms of subsets. To approximate the posterior predictive distribution under the aforementioned conditions, they combine local estimates, divided by the GP prior. Looking to our particular *recyclable* framework, this strategy is analogous to the one presented in Eq. (3.15), but in the logarithmic plane instead and the variational inference setup.

3.1.3 Capabilities of Recyclable Gaussian Processes

We highlight several use cases for the proposed framework. The idea of recycling GP models opens the door to multiple extensions, with a particular attention to the human behavior learning problem. That is, the local-global modelling of heterogeneous data and the adaptation of model complexity in a data-driven manner.

Global Prediction – Our purpose might be to predict how likely an output test datum y_t is at some point $\boldsymbol{x}_t$ of the input space $\mathbb{R}^p$. In this case, the global predictive distribution can be approximated as $p(y_t|\mathcal{D}) \approx \int p(y_t|\boldsymbol{f}_t)q(\boldsymbol{f}_t)d\boldsymbol{f}_t$, with the variational (output function) distribution $q(\boldsymbol{f}_t)$ coming from the integration $q(\boldsymbol{f}_t) = \int p(\boldsymbol{f}_t|\boldsymbol{u}_*)q(\boldsymbol{u}_*)d\boldsymbol{u}_*$. As mentioned in the previous subsection, the integral can also be obtained by quadratures when the solution is intractable.

Recyclable GP and new data – In practice, it might not be necessary to distribute the whole dataset $\mathcal{D}$ in parallel tasks or nodes, with only a few subsets of data $\mathcal{D}_k$ available for the global ensemble. Instead, it is possible to combine the samples in $\mathcal{D}_k$ with the dictionary of local GP variational distributions. In such cases, we would only approximate the likelihood terms in Eq. (3.7) related to the distributed subsets of samples. The resulting *combined* bound between local variational approximations and new unseen data would be equivalent to Eq. (3.15) with additional expectation terms on the new output observations.

Stationarity and expresiveness – We typically assume that the non-linear function $f(\cdot)$ is stationary accross subsets of data. If this assumption is relaxed, some form of adaptation or *forgetting* should be included to match the local GP models. Other types of methods can be also considered for the ensemble, as for instance, with several latent functions (Lázaro-Gredilla and Titsias, 2011) or sparse multi-output GPs (Álvarez and Lawrence, 2011). The current framework also accepts GPs with increased expressiveness. For example, to get multi-modal likelihoods, we can use mixtures of GP experts (Rasmussen and Ghahramani, 2002).

Data-driven complexity and recyclable ensembles – One of the main advantages of the recyclable GP framework is that is allows to perform *data-driven* updates of the model complexity. That is, if an ensemble ends in a new variational GP model, it also can be recycled. Hence, the number of global inducing-variables M can be iteratively increased conditioned to the amount of samples considered. A similar idea was already commented as a potential application of the sparse order-selection theorem by Burt et al. (2019).

MODEL RECYCLING AND USE CASES – The ability of recycling GP models in future or additional global tasks has a significant impact the human behavior applications, where fitted private-owned models in smartphones can be shared for external predictions rather than the data itself. Its application to medicine is also of high interest for us and particularly, for this thesis. If one has a personalized GP model for every patient, epidemiologic surveys can be built from them without centralising private data. This is, perhaps, the key contribution of the *recyclable* GP framework of this section.

3.1.4 Related Work on Distributed Gaussian Processes

The flexible nature of GP models for defining prior densities over non-linear functional spaces has made them a suitable alternative in many probabilistic regression and classification problems. However, GP models are not immune to settings where the model itself needs to adapt to *irregular* ways of accessing the data, e.g. asynchronous observations or missings input areas. Such settings, together with the GP's well-known computational cost for the exact solutions has motivated plenty of approaches focused on *parallelising inference*.

Regarding the task of distributing the computational load between learning *agents* or *nodes*, several GP models have been inspired by the *local experts* Jacobs et al. (1991); Hinton (2002). Mainly, two seminal GP works exploited this connection before the modern era of sparse variational approximations. While the Bayesian committee machine (BCM) of Tresp (2000) focused on merging independently trained Gaussian processes on subsets of the same data, the infinite mixture of GP experts (Rasmussen and Ghahramani, 2002) increased the model expresiveness by combining local GP experts. The approach presented in Sec. 3.1 is closer to the first method, whilst the second one is also amenable but out of the spirit of this thesis.

The emergence of large datasets, with size $N > 10^4$, led to the introduction of approximate models, that in combination with variational inference (Titsias, 2009a), succeed in scaling up GPs. Two more recent approaches that combine sparse GPs with ideas from distributed models or computations are Gal et al. (2014) and Deisenroth and Ng (2015). Based on the variational GP scheme of Titsias (2009a), Gal et al. (2014) presented a novel reparameterisation of the lower bound. This approach allows to distribute the computational cost accross several nodes, being also applicable to GPs with stochastic variational inference (Hensman et al., 2013a) and with non-Gaussian likelihoods Hensman et al. (2015a); Saul et al. (2016).

Out of the sparse approximations to GPs, more inspired in Tresp (2000) and product of experts (Bordley, 1982), the distributed GP model of Deisenroth and Ng (2015) scaled up the parallelisation mechanism of local experts to the range of $N{>}10^6$. Their approach is focused on exact GP regression, not considering classification or other non-Gaussian likelihoods. In Tab. 3.1, we provide a description of the different methods and their main properties, also if each distributed node is a GP itself or not.

Additionally, if we look to the property of having nodes that contain *usable* GP models showed in Tab. 3.1, the present approach is similar to Deisenroth and Ng (2015); Cao and Fleet (2014) and Tresp (2000). The main difference is that we introduce variational approximation methods for non-Gaussian likelihoods. Another detail is that the idea of exploiting properties of full (infinite-dimensional) stochastic processes Matthews et al. (2016) for substituting likelihood probabilities in a general bound has been previously explored in Bui et al. (2017a) and Moreno-Muñoz et al. (2018). Whilst the preliminary work in Bui et al. (2017a) ends in the derivation of expectation-propagation (EP) methods for streaming inference in GPs, the introduction of the distributed reparameterisation of Gal et al. (2014) maker our inference more natural.

Table 3.1: Main properties of distributed GP models in the literature

MODEL	$\mathcal{N}$ REG.	non-$\mathcal{N}$ REG.	CLASS.	HET.	INFERENCE	$\mathcal{GP}_{\text{NODE}}$	DATA ST.
Tresp (2000)	✓	✗	✗	✗	Analytical	✓	✓
Ng and Deisenroth (2014)	✓	✗	✗	✗	Analytical	✓	✓
Cao and Fleet (2014)	✓	✗	✗	✗	Analytical	✓	✓
Deisenroth and Ng (2015)	✓	✗	✗	✗	Analytical	✓	✓
Gal et al. (2014)	✓	✓	✓	✗	Variational	✗	✗
Moreno-Muñoz et al. (2020a)	✓	✓	✓	✓	Variational	✓	✗

(*) Respectively, Gaussian and non-Gaussian regression ($\mathcal{N}$ & non-$\mathcal{N}$ REG), classification (CLASS), heterogeneous (HET) and storage (ST).

There is also the inference framework of Bui et al. (2018) for both federated and continual learning scenarios, but focused on EP and the Bayesian approach of Nguyen et al. (2018). A short analysis of its applications to GPs is included for continual learning settings, but far from the large-scale scope of the method presented in this thesis. Moreover, the spirit of using inducing-points as pseudo-approximations of local subsets of data is shared with Bui and Turner (2014), that comments its potential use for distributed setups. More oriented to dynamical modular models, we find the work by Velychko et al. (2018), whose factorisation across tasks is similar to Ng and Deisenroth (2014) but oriented to state-space models.

3.2 Continual Multi-task Gaussian Processes

The resurgence of interest on probabilistic adaptative methods shows us that, the better the models are adapted to time evolving behaviors, the easier its applicability on real-world problems is. A remarkable evidence of *how necessary* this real-time adaptation is for machine learning can be deduced from multiple applications, e.g. systems for intensive care units (ICU) patients or electronic health records (EHR). In the context of this thesis, it is a key milestone for the purpose of modelling human behavior in an *online* manner, and as we will see in the next chapter, having adaptive probabilistic models to heterogeneous high-dimensional data is the first step forward to the detection of anomalous events or outliers.

Among all the adaptive approaches that can be considered, in this thesis and particularly in this chapter, we focus on the *continual* methods. Continual learning, also known as life-long learning, is a very general family of *online* learning methods whose principal properties are the adaptation to non i.i.d. data, characterization of tasks that evolve over time and capture of new emergent patterns previously unseen by the model itself.

3.2.1 Continual Gaussian Processes

Gaussian process models are not excluded from the need of real-time adaptation. Despite their extended use in temporal applications, recursively updating their parameters without *revisiting* training samples is not a trivial problem yet. Particularly in GPs, the difficulties are double. First, the estimation of non-linear latent functions is constrained by the same principles of *online* Bayesian learning, that is, how to re-introduce former posterior discoveries as new prior beliefs. Second, due to GP priors are based on the construction of covariance matrices via kernel functions, incrementally adapting such matrices to new incoming samples requires expensive ways of matrix completion or even unfeasible inversions when large data is observed.

However, there has been a noticeable effort on adapting GP models for sequential input-output observations over the past decades. As standard Gaussian GP regression scenarios

are usually accompanied by tractable solutions, preliminary works focused exclusively on the iterative counterpart. In particular, this paradigm attracted significant attention since the seminal works by Csató and Opper (2002) and Girard et al. (2003) presented the two preliminar alternatives to perform online predictions. One used *moment matching* to fit sequential posterior distributions from one single recent sample. In the second case, motivated by one-step ahead predictions, they incorporated an additive input in an equivalent state-space model, which consists of mappings over the last few observed instances, L steps back.

Besides initial approaches to *online* GPs, other recent works have also addressed the continual learning problem. For example, sequential rank-one updates of locally trained GPs were proposed in Nguyen-Tuong et al. (2008), where they also introduce an inclusion-deletion strategy of data points for the model online adaptation. The GP is learned by EP as in Henao and Winther (2010). Also for the single-output GP case, but closer to the scalable framework developed in this thesis, we find that the stochastic gradient descent method in Hensman et al. (2013a) for Gaussian regression and Hensman et al. (2015a) for classification, are applicable to online settings. However, it would require to consider ever-increasing datasets, which *a priori* might be problematic. Another recent example is the semi-described (missing inputs) and semi-supervised (missing outputs) GP learning model in Damianou and Lawrence (2015). Here, forecasting regression problems are considered as a semi-described scenarios where predictions are obtained iteratively in an *auto-regressive* way.

In terms of scalability for single-output GP models, both Cheng and Boots (2016) and Bui et al. (2017a) extended online learning methods and uncertainty propagation to the popular variational inference setup of sparse GP approximations. They used a novel KL divergence formulation that constrains the new fitted distribution w.r.t. the one in the previous instant. While the first work is only related to univariate Gaussian regression problems, the last approach has the additional advantage of accepting limited non-Gaussian likelihoods as well as it is able to include α-divergences for more general inference. This last method is analysed in Nguyen et al. (2017) within its theoretical bounds.

An exception to the previous works is the approach in Solin et al. (2018). Instead of employing sparse methods, the authors use the approximate Markovian structure of Gaussian processes to reformulate the problem as a state-space model. Within this framework, the complexity is reduced from cubic to linear cost in the number of observations, but still stays unfeasible w.r.t. the number of states. Introducing a fast EP scheme helps to overcome the issue. Additionally, the model is capable to perform the online learning of kernel hyperparameters as well as dealing with non-Gaussian likelihoods.

Sequential Data Formulation

Consider supervised learning scenarios where pairs of input-output data $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^N$ are observed in a *sequential* manner, with $\boldsymbol{x} \in \mathbb{R}^p$ as in the previous GP approaches. Output targets y_n can be either continuous or discrete. We assume the sequential observation process to be a finite stream of smaller subsets or batches, such that $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_T\}$. Additionally, each $t-$th batch, $\mathcal{D}_t = \{\boldsymbol{x}_n, y_n\}_{n=1}^{N_t}$, may have an irregular size, this is, different length per data partition and $N_t < N$ i all cases. This data notation is analogous to the one presented in the previous Sec. 3.1.

From the GP perspective, we also consider that every output sample is generated as $y_n \sim p(y_n|f_n)$, where f_n is the non-linear function $f(\boldsymbol{x}_n)$. As we assumed in Sec. 3.1, the latent process parameterizes the likelihood model and it is drawn from a GP prior $f \sim \mathcal{GP}(0, k(\cdot,\cdot))$, where $k(\cdot,\cdot)$ can be any valid covariance function. The zero-mean is assumed for simplicity in the derivations.

Since we do not know when the next subset $\mathcal{D}_t$ arrives at each time-step $t-1$, the waiting time and memory allocation resources cannot be estimated *a priori*. Commonly, due to the size of the batches is irregular and also unknown. Based on Bui et al. (2017a), we assume that receiving the entire sequence of data and computing the posterior distribution $p(f|\mathcal{D})$ is unfeasible and extremely high-time demanding. As alternative, we will consider a continual learning strategy, which refers to the ability of adapting models in an *online* fashion when data samples are not i.i.d. updating their parameters without re-observing the entire data sequence.

In what follows, we will use the notation $\mathcal{D} = \{\mathcal{D}_{\text{old}}, \mathcal{D}_{\text{new}}\}$, where the collection $\mathcal{D}_{\text{old}} = \{\boldsymbol{x}_{\text{old}}, \boldsymbol{y}_{\text{old}}\}$ makes reference to all variables seen so far and the collection $\mathcal{D}_{\text{new}} = \{\boldsymbol{x}_{\text{new}}, \boldsymbol{y}_{\text{new}}\}$ is defined as the smaller subset of *new* incoming variables. For this construction, notice that if $\mathcal{D}_t$ is observed at a given time t, the *old* collection would correspond to $\mathcal{D}_{\text{old}} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_{t-1}\}$, while $\mathcal{D}_{\text{new}} = \mathcal{D}_t$. This results in an ever-increasing collection $\mathcal{D}_{\text{old}}$ that is recursively evaluated for fitting the GP model.

Sparse Approximations for Sequences

Exact inference in the standard GP is widely known for its $\mathcal{O}(N^3)$ complexity for training and $\mathcal{O}(N^2)$ per test prediction as we showed in Ch. 2. In our sequential case, given the previous description of the *new* and *old* data collections, the computational cost for learning the exact GP function could be even more intensive, with a recurrent complexity of $\mathcal{O}(N_1^3), \mathcal{O}((N_1 + N_2)^3), \ldots, \mathcal{O}(N^3)$ for prediction. Following the same approach as in Ch. 2 for heterogeneous MOGPs and, as in Sec. 3.1 for parallelisable GPs, we introduce inducing inputs (Snelson and Ghahramani, 2006).

Importantly, we follow the same variational derivation of Titsias (2009a), previously introduced in this thesis. Here, the choice for the joint auxiliary density $q(\boldsymbol{f}, \boldsymbol{u})$ is to factorize according to $q(\boldsymbol{f}, \boldsymbol{u}) \approx p(\boldsymbol{f}|\boldsymbol{u})q(\boldsymbol{u})$. This reduces the variational inference problem to learn the parameters of the distribution $q(\boldsymbol{u})$, that we also assume to be Gaussian.

As the starting point for our continual derivation, we condition every observed output y_n to the augmented infinite-dimensional GP, that is f_∞, as we did in Sec. 3.1 and also based on Bui et al. (2017a). This leads to have likelihood densities $p(y_n|f_\infty)$ instead. Every infinite-dimensional variable f_∞ contains all the instances of the process $f(\cdot)$. This also includes the inputs $\{\boldsymbol{x}_n\}_{n=1}^N$ and the inducing variables $\{\boldsymbol{z}_m\}_{m=1}^M$. The idea play a key role in the development of the continual update, as it did in the mechanism for the distributed GPs.

From the perspective of two collections, one being *old* and another formed by *new* observations, if we introduce the augmented likelihood model $p(y_n|f_\infty)$ induced by f_∞, the log-marginal density over the sequential data turns to be an infinite-dimensional integral. This can be decomposed as

$$\log p(\boldsymbol{y}_{\text{old}}, \boldsymbol{y}_{\text{new}}) = \log \int p(\boldsymbol{y}_{\text{old}}, \boldsymbol{y}_{\text{new}}|f_\infty)p(f_\infty)df_\infty, \tag{3.21}$$

where $p(f_\infty)$ corresponds to the *full* GP prior over the stochastic process. That is, we place a density over all instances of $f(\cdot)$ in the input space $\mathbb{R}^p$. We suppose now, that both output data collections $\boldsymbol{y}_{\text{old}}$ and $\boldsymbol{y}_{\text{new}}$ are non i.i.d. but conditioned to the whole process f_∞. This assumption allows us to apply *conditional independence*. This leads us to obtain a factorized likelihood function of the form

$$p(\boldsymbol{y}_{\text{old}}, \boldsymbol{y}_{\text{new}}|f_\infty) = p(\boldsymbol{y}_{\text{old}}|f_\infty)p(\boldsymbol{y}_{\text{new}}|f_\infty), \tag{3.22}$$

as is considered in Bui et al. (2017a) and Moreno-Muñoz et al. (2020a). We remark that both terms in the likelihood factorization are now the *augmented* version of the densities, and

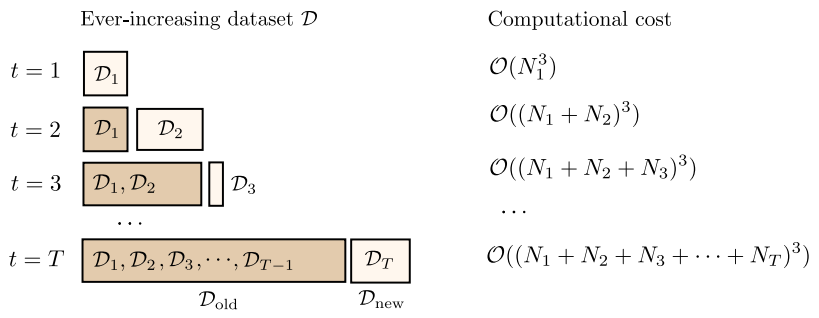


Figure 3.2: Illustration of the *ever*-increasing sequential dataset $\mathcal{D}$ and the relative computational cost metrics conditioned to the number of observations. New observed batches $\mathcal{D}_t$ are of irregular size that is unknown *a priori*. Darker boxes indicate the $\mathcal{D}_{\text{old}}$ subset at each time step, and the clearer counterpart are $\mathcal{D}_{\text{new}}$. The size of boxes indicates the number of samples in each subset. The computational cost of the *exact* GP inference solution is indicated in the right column.

also conditional independent given f_∞. This idea is also used for modelling the distributed partitions in the model of Sec. 3.1.

Then, any standard lower bound $\mathcal{L}$, built from the log-marginal expression in Eq. (3.21), would require to evaluate expectations of the form $\mathbb{E}_{q(f_\infty)}[\log p(\boldsymbol{y}_{\text{old}}, \boldsymbol{y}_{\text{new}}|f_\infty)]$, where the variational density is obtained from $q(f_\infty) = \int p(f_\infty|\boldsymbol{u})q(\boldsymbol{u})d\boldsymbol{u}$ as in the uncollapsed versions of the bound (Lázaro-Gredilla and Titsias, 2011; Hensman et al., 2012). Having the previously introduced factorization between the *old* and *new* variables, the expectation term of the bound can be divided in two, for example, $\mathbb{E}_{q(f_\infty)}[\log p(\boldsymbol{y}_{\text{old}}|f_\infty)]$ and $\mathbb{E}_{q(f_\infty)}[\log p(\boldsymbol{y}_{\text{new}}|f_\infty)]$. Notice that the main problem comes from the evaluation of these two terms, as the computation of the lower bound is extremely *unbalanced*. Mainly, due to the difference of size between $\boldsymbol{y}_{\text{old}}$ and $\boldsymbol{y}_{\text{new}}$ might be huge, e.g. millions of samples vs. hundreds respectively. This fact results in very long time computation for re-training any variational GP model with a few more recent observations included in the ever-increasing dataset (see Fig. 3.2), due to the dimensionality of the likelihood density $p(\boldsymbol{y}_{\text{old}}|f_\infty)$.

To circumvent this issue, we investigate potential ways for avoiding the sequential revisiting of old densities $p(\boldsymbol{y}_{\text{old}}|f_\infty)$, or at least, approximating them with lower computational cost. In the next sections, this goal is achieved for both single-output and multi-output GP models.

Recurrent Prior Reconstruction

A meaningful solution for avoiding the sequential evaluation of ever-increasing datasets is approximating their augmented *old* likelihood densities $p(\boldsymbol{y}_{\text{old}}|f_\infty)$ using the previously inferred (joint) variational distribution $q(f_\infty|\boldsymbol{\phi}_{\text{old}})$ at each time-step. Here, the infinite-dimensional notation of the variational distribution is just momentary, as we will later apply properties of Gaussian marginals to reduce its dimensionality to the corresponding function instances of interest.

The idea of approximating the old likelihood densities to avoid their re-visiting was first introduced in Bui et al. (2017a), based on the similar strategy carried out within expectation-propagation (EP) inference. Thus, if we apply the Bayes rule theorem, the variational posterior distribution can be approximated as

$$q(f_\infty|\boldsymbol{\phi}_{\text{old}}) \approx p(f_\infty|\boldsymbol{y}_{\text{old}}, \boldsymbol{x}_{\text{old}}) \propto p(f_\infty)p(\boldsymbol{y}_{\text{old}}|f_\infty), \tag{3.23}$$

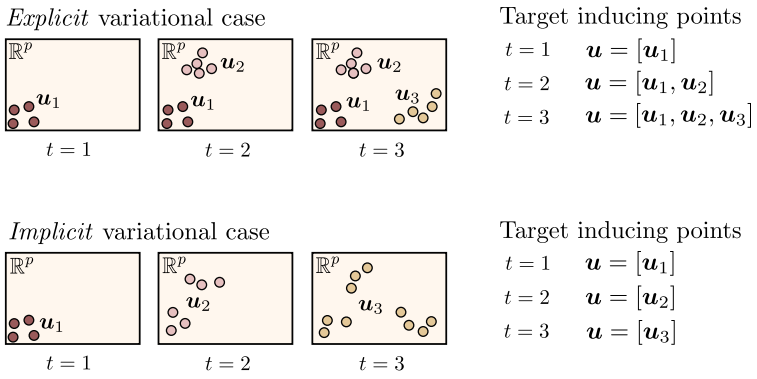


Figure 3.3: Illustration of the input space $\mathbb{R}^p$ in the two studied cases for treating the inducing points $\boldsymbol{u}$ in a continual learning manner. In the *explicit* case, the past inducing points $\boldsymbol{u}_{t-1}$ are fixed in all time-steps t, and must be concatenated with the new ones $\boldsymbol{u}_t$. This makes the model to increase the computational complexity M in every iteration. In the *implicit* case, the past inducing points $\boldsymbol{u}_{t-1}$ are integrated out, and the new ones $\boldsymbol{u}_t$ explore other regions of the input space as well as capture the uncertainty in the already learned areas.

where the equality can be inverted to give us a proportional estimate of the form

$$p(\boldsymbol{y}_{\text{old}}|f_\infty) \approx \frac{q(f_\infty|\boldsymbol{\phi}_{\text{old}})}{p(f_\infty)}. \tag{3.24}$$

Having this recursive approximation in Eq. (3.24) for old likelihoods conditioned on the past parameters $\boldsymbol{\phi}_{\text{old}}$, we can use it to build a lower bound $\mathcal{L}$ where the data re-visiting is avoided.

Under this strategy, the variational distribution $q(f_\infty|\boldsymbol{\phi}_{\text{old}})$ factorises according to

$$q(f_\infty|\boldsymbol{\phi}_{\text{old}}) = p(f_{\infty\neq\boldsymbol{u}}|\boldsymbol{u},\boldsymbol{\phi}_{\text{old}})q(\boldsymbol{u}|\boldsymbol{\phi}_{\text{old}}), \tag{3.25}$$

where $f_\infty = \{f_{\infty\neq\boldsymbol{u}} \cup \boldsymbol{u}\}$ as in the previous Sec. 3.1. The main problem that we encounter here is on re-using the distributions $q(\boldsymbol{u}|\boldsymbol{\phi}_{\text{old}})$ estimated over a (now) *fixed* number of inducing inputs $\boldsymbol{Z}_{\text{old}}$. That is, the subset of inducing inputs $\boldsymbol{Z}_{\text{old}} = \{\boldsymbol{z}_m\}_{m=1}^{M_{\text{old}}}$ used in the variational inference run of the previous t-th time-step. This is the principal limiting property of the model in Bui et al. (2017a), since a continual learning scenario does not fit well with an ever-fixed distribution over a subset of input points $\boldsymbol{Z}_{\text{old}}$ that we cannot update neither.

The particular re-visiting issue can be understood in terms of the *evolution-exploration* duple of the input space $\mathbb{R}^p$, where a GP model is initially fitted to some region of interest that later on, is not needed anymore. For instance, if we directly plug in $q(\boldsymbol{u}|\boldsymbol{\phi}_{\text{old}})$ using the expression in Eq. (3.24), that we here refer as the *explicit* distribution, then the new variational inference problem will carry out with such a fixed distribution and their corresponding pseudo observations $\boldsymbol{u}$ step-by-step. Here, we say that it is fixed due to posterior uncertainty metrics cannot be modified anymore, i.e. variational parameters $\boldsymbol{\mu}_{\boldsymbol{u}}$ and $\mathbf{S}_{\boldsymbol{u}}$ are fixed, as the old data that we observe is no longer available.

Looking to the inducing points $\boldsymbol{u}$ and their relative parameters $\boldsymbol{\phi}_{\text{old}}$, what we are doing is to recurrently introduce a summary of our data. In terms of a rigorous continual learning approach, this is another way of revisiting past observed samples and it also forces the GP model to *concatenate* old and new subsets $\boldsymbol{u}$, which in the long term could be problematic. Importantly, this point is undesired for certain tasks, i.e. high-dimensional input problems, or due to computational complexity reasons. An illustration of this *explicit* variational problem and the concatenation of subsets of inducing points is shown in Fig. 3.3.

In this section, we introduce the proposed solution, that we also refer as the use of *implicit* variational distributions for the continual learning problem in GP models.

CONTINUAL GAUSSIAN PROCESS PRIOR – Inspired on online Bayesian inference methods, where *past* posterior distributions are usually taken as *future* priors, our main goal is reconstruct the GP prior conditioned on the given parameters $\boldsymbol{\phi}_{\text{old}}$. However, we avoid to treat the last subset of inducing points $\boldsymbol{u}$ explicitly.

The particular construction is as follows. We take the posterior predictive distribution from standard GP models. It is usualliy obtained by direct marginalisation of the posterior probabilities $p(f_\infty|\mathcal{D})$ given the conditional distribution at test inputs $p(\boldsymbol{f}_*|f_\infty)$, whose output values $\boldsymbol{y}_*$ we aim to predict. Typically and under the presence of sparse GP approximations, the predictive distribution integral takes the form

$$p(\boldsymbol{f}_*|\mathcal{D}) = \int p(\boldsymbol{f}_*|\boldsymbol{u})p(\boldsymbol{u}|\mathcal{D})d\boldsymbol{u}. \tag{3.26}$$

In our solution, this posterior predictive formulation is the key idea for recurrently building *continual* GP priors, that is, a new *implicit* distribution at each time-step t, where only the past estimated parameters $\boldsymbol{\phi}_{\text{old}}$ intervene. For its derivation we take the appendix A.2 of Álvarez et al. (2009) as our starting point. Thus, we have a conditional prior of the form

$$p(\boldsymbol{u}_*|\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u}_*|k_{*\boldsymbol{u}}\mathbf{K}_{\boldsymbol{uu}}^{-1}\boldsymbol{u}, k_{**} - k_{*\boldsymbol{u}}\mathbf{K}_{\boldsymbol{uu}}^{-1}k_{*\boldsymbol{u}}^\top), \tag{3.27}$$

• This is the main utility of the presented approach, since we can iteratively place $\boldsymbol{Z}_*$ in the areas of the input space that become relevant as more data is observed in a continual learning setting.

where $\boldsymbol{u}_*$ refers to the function evaluations of $f(\cdot)$ on any arbitrary input-vector $\boldsymbol{Z}_*$ that we may consider. Here, the covariance matrix corresponds to $\mathbf{K}_{\boldsymbol{uu}} \in \mathbb{R}^{M\times M}$, with entries $k(\boldsymbol{z}_i, \boldsymbol{z}_j)$ as $\boldsymbol{z}_i, \boldsymbol{z}_j \in \boldsymbol{Z}_{\text{old}}$ and $k_{*\boldsymbol{u}} = [k(\cdot, \boldsymbol{z}_1), \cdots, k(\cdot, \boldsymbol{z}_M)]^\top$. In a similar manner, $k_{**} = k(\cdot, \cdot)$ as in the kernel function of any GP prior.

Having the conditional distribution in Eq. (3.27), we see that it combines covariance matrices from both explicit and implicit distributions, e.g. $\mathbf{K}_{\boldsymbol{uu}}$ and k_{**}, respectively. Then, based in the posterior predictive integral in Eq. (3.26), we can extend the expectation operator to accept the variational distribution $q(\boldsymbol{u}|\boldsymbol{\phi}_{\text{old}})$ instead of $p(\boldsymbol{u}|\mathcal{D})$. This makes the conditional GP prior $p(u_*|\boldsymbol{u})$ behave as the former approximate posterior density q indicates. The whole process results in a novel *continual* distribution, formally denoted as $\widetilde{q}(u_*|\boldsymbol{\phi}_{\text{old}})$, that we obtain as

$$\widetilde{q}(u_*|\boldsymbol{\phi}_{\text{old}}) \approx \int p(u_*|\boldsymbol{u})q(\boldsymbol{u}|\boldsymbol{\phi}_{\text{old}})d\boldsymbol{u}, \tag{3.28}$$

where we remark that variables u_* are not fixed yet. Additionally, if we assume that the variational distribution $q(\boldsymbol{u}|\boldsymbol{\phi}_{\text{old}}) = \mathcal{N}(\boldsymbol{u}|\boldsymbol{\mu}_{\text{old}}, \mathbf{S}_{\text{old}})$, then our parameters $\boldsymbol{\phi}_{\text{old}}$ become $\boldsymbol{\phi}_{\text{old}} = \{\boldsymbol{\mu}_{\text{old}}, \mathbf{S}_{\text{old}}\}$. Then, the previous expression in Eq. (3.28) leads us to a tractable integral over Gaussian distributions that results in the closed-form formula of the continual GP prior. Its form is

$$u_* \sim \mathcal{GP}(k_{*\boldsymbol{u}}\mathbf{K}_{\boldsymbol{uu}}^{-1}\boldsymbol{\mu}_{\text{old}}, k_{**} + k_{*\boldsymbol{u}}\mathbf{K}_{\boldsymbol{uu}}^{-1}(\mathbf{S}_{\text{old}} - \mathbf{K}_{\boldsymbol{uu}})\mathbf{K}_{\boldsymbol{uu}}^{-1}k_{*\boldsymbol{u}}^\top). \tag{3.29}$$

A similar expression was derived in Burt et al. (2019), where a theoretical analysis on sparse GP regression is performed out of the continual learning problem. This same framework was later extended in Burt et al. (2020). In our particular case, the conditional GP prior in Eq. (3.29) coincides with the approximated posterior process that VI on sparse GP models aims to minimize through the KL divergence (Matthews et al., 2016).

This result is of particular interest to us, since it provides a closed-form way to introduce Bayesian *online* learning into GP models, allowing us to naturally avoid any data revisiting; only passing past parameters forward and fixing the posterior-prior recursion of our framework.

Continual Lower Bounds

Exact posterior inference in the continual setting is still intractable using the previous framework and variational methods are still required. However, we are now able to sequentially build lower bounds, e.g. one bound below $p(\boldsymbol{y})$ per time-step, by only updating from a few recent observations $\mathcal{D}_{\text{new}}$.

This turns the continual problem to be a recursive process where a new variational tool appears at each iteration and inherits the previous learned uncertainty metrics. We determine these *continual lower bounds* as $\mathcal{L}_{\mathcal{C}}$, and are obtained as follows

$$\log p(\boldsymbol{y}_{\text{new}}, \boldsymbol{y}_{\text{old}}) \leq \mathcal{L}_{\mathcal{C}} \approx \int q(f_\infty|\boldsymbol{\phi}_{\text{new}}) \log \frac{p(\boldsymbol{y}_{\text{new}}|f_\infty) q(f_\infty|\boldsymbol{\phi}_{\text{old}}) p(f_\infty|\boldsymbol{\psi}_{\text{new}})}{q(f_\infty|\boldsymbol{\phi}_{\text{new}}) p(f_\infty|\boldsymbol{\psi}_{\text{old}})} df_\infty, \quad (3.30)$$

where $q(f_\infty|\boldsymbol{\phi}_{\text{new}})$ is the new *augmented* variational distribution that we want to recursively update, and $\boldsymbol{\psi}_{\text{old}}$ and $\boldsymbol{\psi}_{\text{new}}$ are the past and current subsets of hyperparameters of the GP prior, respectively. Notice that the distribution $q(f_\infty|\boldsymbol{\phi}_{\text{new}})$ may also factorise according to $p(f_{\infty \neq \boldsymbol{u}_*}|\boldsymbol{u}_*, \boldsymbol{\phi}_{\text{new}}) q(\boldsymbol{u}_*|\boldsymbol{\phi}_{\text{new}})$. We often use this $\boldsymbol{\psi}$ to refer both $\boldsymbol{\psi}_{\text{old}}$ and $\boldsymbol{\psi}_{\text{new}}$ simultaneously, i.e. $\boldsymbol{\psi} = \{\boldsymbol{\psi}_{\text{old}}, \boldsymbol{\psi}_{\text{new}}\}$.

• In the *vanilla* GP case with a Gaussian likelihoo model and RBF kernel, the hyperparameters $\boldsymbol{\psi}$ would correspond to the amplitude σ_a, the lengthscale ℓ and the observation noise σ_n.

Again, to avoid the data revisiting, we have substituted the past likelihood term $p(\boldsymbol{y}_{\text{old}}|f_\infty)$ by its unnormalised approximation, taken from the inverted Bayes rule in Eq. (3.24). A key difference with respect to Bui et al. (2017a) appears on the factorisation of our past variational distribution $q(f_\infty|\boldsymbol{\phi}_{\text{old}})$. Now, instead of conditioning on a fixed number of inducing points $\boldsymbol{u}$, we make use of the continual GP prior in Eq. (3.29), leading to

$$q(f_\infty|\boldsymbol{\phi}_{\text{old}}) = p(f_{\infty \neq \boldsymbol{u}_*}|\boldsymbol{u}_*, \boldsymbol{\psi}_{\text{old}}) \widetilde{q}(\boldsymbol{u}_*|\boldsymbol{\phi}_{\text{old}}), \quad (3.31)$$

where we extended the factorisation of Titsias (2009a) to accept the augmented infinite-dimensional function space f_∞. Moreover, it now makes sense to reduce the lower bound $\mathcal{L}_{\mathcal{C}}$ in Eq. (3.30) by critically canceling all conditionals of the form $p(f_{\infty \neq \boldsymbol{u}_*}|\boldsymbol{u}_*)$. Notice that we use $f_\infty = \{f_{\infty \neq \boldsymbol{u}_*} \cup \boldsymbol{u}_*\}$ to apply CI all over the augmented distributions. Further details are provided in the appendix of this thesis. Then, we are able to obtain a *triple*-termed bound

$$\begin{aligned} \mathcal{L}_{\mathcal{C}} = \int q(f_\infty|\boldsymbol{\phi}_{\text{new}}) \log p(\boldsymbol{y}_{\text{new}}|f_\infty) df_\infty - \int q(f_\infty|\boldsymbol{\phi}_{\text{new}}) \log \frac{q(\boldsymbol{u}_*|\boldsymbol{\phi}_{\text{new}})}{p(u_*|\boldsymbol{\psi}_{\text{new}})} df_\infty \\ + \int q(f_\infty|\boldsymbol{\phi}_{\text{new}}) \log \frac{\widetilde{q}(u_*|\boldsymbol{\phi}_{\text{old}})}{p(u_*|\boldsymbol{\psi}_{\text{old}})} df_\infty. \end{aligned} \quad (3.32)$$

Having this augmented version of $\mathcal{L}_{\mathcal{C}}$, we are now interested in the derivation of a closed-form expression that can be evaluated on a specific number of inducing inputs $\boldsymbol{Z}$ rather than on the infinite-dimensional integrals w.r.t. f_∞.

For this purpose, suppose that our new incoming samples $\mathcal{D}_{\text{new}}$ contain a subset of input values $\boldsymbol{x}_{\text{new}}$ whose distance from all the previous ones $\boldsymbol{x}_{\text{old}}$ is significant. It makes sense to increase the capacity of $\boldsymbol{Z}$ in order to refine the approximate posterior (Burt et al., 2019). As a consequence, we introduce a new set of inducing variables $\boldsymbol{Z}_{\text{new}} = \{\boldsymbol{z}_m\}_{m=1}^{M_{\text{new}}}$, where the vector $\boldsymbol{u}_{\text{new}}$ of function evaluations corresponds to $\boldsymbol{u}_{\text{new}} = [u(\boldsymbol{z}_1), \cdots, u(\boldsymbol{z}_{M_{\text{new}}})]^\top$. Notice that we now aim to update the distribution $q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}}) = \mathcal{N}(\boldsymbol{u}_{\text{new}}|\boldsymbol{\mu}_{\text{new}}, \mathbf{S}_{\text{new}})$ where $\boldsymbol{\phi}_{\text{new}} = \{\boldsymbol{\mu}_{\text{new}}, \mathbf{S}_{\text{new}}\}$ in this particular case.

One strategy to introduce the previous subset is that all the distributions that make reference to the predictive variables $\boldsymbol{u}_*$ in $\mathcal{L}_{\mathcal{C}}$ are substituted by $\boldsymbol{u}_{\text{new}}$. That is, the former prediction at test-inputs $\boldsymbol{Z}_*$ used in the *implicit* distribution formalism in Eq. (3.29) is now computed at $\boldsymbol{Z}_{\text{new}}$.

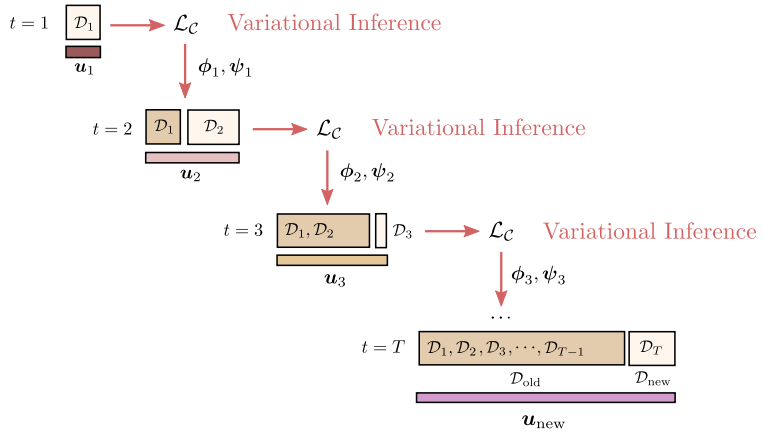


Figure 3.4: Diagram of the continual GP learning mechanism. Brown-colored boxes with $\mathcal{D}_t$ inside indicate the observation of a new batch of samples. Lower rectangles in colors represent the inducing points $\boldsymbol{u}_t$ used during each iteration. The letter $\mathcal{L}_\mathcal{C}$ indicates the step where a new continual lower bound is maximised using variational inference methods. Downside red arrows with $\boldsymbol{\phi}, \boldsymbol{\psi}$, variational parameters and hyperparameters respectively, indicate the single uncertainty metrics passed to the next iteration.

In addition, except for the log-likelihood term in Eq. (3.32), distributions on f_∞ may factorise, for instance as $q(f_\infty|\boldsymbol{\phi}_{\text{new}}) = q(f_{\infty\neq\boldsymbol{u}_{\text{new}}}|\boldsymbol{u}_{\text{new}}, \boldsymbol{\psi}_{\text{new}})p(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})$, particularly the variational ones. This convenient factorisation allows us to apply properties of Gaussian marginals, integrating all function values $f_{\infty\neq\boldsymbol{u}_{\text{new}}}$ out of the $\mathcal{L}_\mathcal{C}$ bound.

Given the reduction of $\mathcal{L}_\mathcal{C}$, we are also able to obtain a closed-form expression of the bound in Eq. (3.32) where three prior and one posterior distributions intervene. Respectively, the key terms that we remark are:

i) the *new* GP prior distribution $p(\boldsymbol{u}_{\text{new}}|\boldsymbol{\psi}_{\text{new}})$,

ii) the *old* GP prior distribution $p(\boldsymbol{u}_{\text{new}}|\boldsymbol{\psi}_{\text{old}})$,

iii) the *continual* GP prior distribution $\widetilde{q}(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{old}})$ and,

iv) the variational posterior distribution $q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})$.

Then, using the previous distributions, we can further simplify $\mathcal{L}_\mathcal{C}$ to be

$$\begin{aligned}\mathcal{L}_\mathcal{C} = \mathbb{E}_{q(\boldsymbol{f}_{\text{new}})}[\log p(\boldsymbol{y}_{\text{new}}|\boldsymbol{f}_{\text{new}})] - \text{KL}[q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})||p(\boldsymbol{u}_{\text{new}}|\boldsymbol{\psi}_{\text{new}})] \\ + \text{KL}[q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})||p(\boldsymbol{u}_{\text{new}}|\boldsymbol{\psi}_{\text{old}})] - \text{KL}[q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})||\widetilde{q}(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{old}})], \quad (3.33)\end{aligned}$$

where $q(\boldsymbol{f}_{\text{new}}) = \int p(\boldsymbol{f}_{\text{new}}|\boldsymbol{u}_{\text{new}})q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})d\boldsymbol{u}_{\text{new}}$ as in Saul et al. (2016), with $\boldsymbol{f}_{\text{new}}$ being the vector of output function evaluations $f(\cdot)$ over the inouts $\boldsymbol{x}_{\text{new}}$. Importantly, the four distributions described above are involved in the bound in Eq. (3.33) via Kullback-Leibler (KL) divergence operators. We identify these KL terms as regularizers forcing the new fitted variational distribution $q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})$ to satisfy two intuitive conditions. First,

• See the analytical expression of $q(\boldsymbol{f}_{\text{new}})$ in the appendix.

to be close enough from the continual GP prior due to $\mathrm{KL}[q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})||\widetilde{q}(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{old}})]$. Second, to be in between the new and old GP prior distributions, via the substraction $\mathrm{KL}[q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})||p(\boldsymbol{u}_{\text{new}}|\boldsymbol{\psi}_{\text{new}})] - \mathrm{KL}[q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})||p(\boldsymbol{u}_{\text{new}}|\boldsymbol{\psi}_{\text{old}})]$. Notice that an extra minus sign is later added to this term in Eq. (3.33).

We know from the contrastive bounds in Ruiz and Titsias (2019) that the substraction of two KL divergences is not a divergence itself. However, having the third additive term, this point changes. Moreover, the functional form of the bound $\mathcal{L}_{\mathcal{C}}$ also simplifies to the continual learning process illustrated in Fig. 3.4, to recurrently make the following update of parameters

$$\boldsymbol{\phi}_{\text{old}}^{(t+1)} \leftarrow \boldsymbol{\phi}_{\text{new}}^{(t+1)} := \arg\max_{\boldsymbol{\phi}_{\text{new}}} \left[\mathcal{L}_{\mathcal{C}} \left(\mathcal{D}_{\text{new}}^{(t)}, \boldsymbol{\phi}_{\text{old}}^{(t)} \right) \right]. \tag{3.34}$$

From a practical point of view, when $t = 0$ in the expression above, that is, in the first step, we train the model using the bound in Hensman et al. (2015a) in order to set $\boldsymbol{\phi}_{\text{new}}^{(0)}$. The complete recursive computation of Eq. (3.33) is detailed in Alg. 1. Moreover, to learn the variational parameters $\boldsymbol{\phi}_{\text{new}} = \{\boldsymbol{\mu}_{\text{new}}, \mathbf{S}_{\text{new}}\}$, we represent the covariance matrix as $\mathbf{S}_{\text{new}} = \mathbf{L}_{\text{new}}\mathbf{L}_{\text{new}}^{\top}$. As a consequence, we maximise $\mathcal{L}_{\mathcal{C}}$ w.r.t. the triangular lower matrix $\mathbf{L}_{\text{new}}$ to ensure positive definiteness when used unconstrained optimization.

Computational cost – In terms of computational cost, the three KL divergence terms in Eq. (3.33) are analytically tractable and of equal dimension, i.e. M_{new}. However, depending on the likelihood model considered for $p(\boldsymbol{y}_{\text{new}}|\boldsymbol{f}_{\text{new}})$, as for instance, Gaussian, Bernoulli or Poisson distributions, the expectations could be intractable. Additionally, if we observe binary samples $y_n \in [0, 1]$, such integrals could be solved via Gaussian-Hermite quadratures, similarly to Hensman et al. (2015a); Saul et al. (2016), as we previously did in Sec. 3.1.

Selection of inducing inputs – The selection of the inducing inputs $\boldsymbol{Z}_{\text{new}}$ is of particular importance for the consistency of the continual learning recursion. Its size, M_{new}, may vary from the number M_{old} of previous inducing-points $\boldsymbol{Z}_{\text{old}}$ without constraints. Notice that, if the incoming batch of samples $\mathcal{D}_t$ is determined by some inputs $\boldsymbol{x}_{\text{new}}$ which explore unseen regions of $\mathbb{R}^p$, then $\boldsymbol{Z}_{\text{new}}$ should be placed to capture this new corresponding area.

Algorithm 1 — Continual Gaussian process learning

1: Initialize $\boldsymbol{\phi}_{\text{new}}^{(0)}$ and $\boldsymbol{\psi}_{\text{new}}^{(0)}$ randomly.
2: **input:** Observe $\mathcal{D}_{\text{new}}^{(0)}$
3: Maximise $\mathcal{L} \leq \log p(\mathcal{D}_{\text{new}}^{(0)})$ w.r.t. $\{\boldsymbol{\phi}_{\text{new}}^{(0)}, \boldsymbol{\psi}_{\text{new}}^{(0)}\}$. // standard variational inference
4: **for** $t \in 1, \ldots, T$ **do**
5: Update $\{\boldsymbol{\phi}_{\text{old}}^{(t)}, \boldsymbol{\psi}_{\text{old}}^{(t)}\} \leftarrow \{\boldsymbol{\phi}_{\text{new}}^{(t-1)}, \boldsymbol{\psi}_{\text{new}}^{(t-1)}\}$ // fitted parameters become the old ones
6: Choose initial $\mathcal{Z}_{\text{new}}$ // initialization of inducing points
7: Compute continual GP prior $\widetilde{q}(\cdot|\boldsymbol{\phi}_{\text{old}}^{(t)})$ // conditional prior reconstruction
8: **input:** Observe $\mathcal{D}_{\text{new}}^{(t)}$
9: Maximise $\mathcal{L}_{\mathcal{C}}$ w.r.t. $\{\boldsymbol{\phi}_{\text{new}}^{(t)}, \boldsymbol{\psi}_{\text{new}}^{(t)}\}$. // continual variational inference
10: **end for**

However, due to we marginalize the former pseudo-observations $\boldsymbol{u}_{\text{old}}$ in Eq. (3.26) for the continual GP prior construction, either $\boldsymbol{Z}_{\text{old}}$ and $\boldsymbol{Z}_{\text{new}}$ are not recommended to coincide on any value. The best practice is to choose robust initializations for $\boldsymbol{Z}_{\text{new}}$. Additional constraints are not needed.

Stochastic Continual Learning

Based on Hensman et al. (2013a), we assume that the likelihood model is conditionally independent and fully factorisable across samples, that is, it holds

$$p(\boldsymbol{y}|\boldsymbol{f}) = \prod_{n=1}^{N} p(y_n|f_n). \tag{3.35}$$

This likelihood factorisation typically leads to the conditional expectations in Eq. (3.33) that are also valid across data observations. This allows us to introduce *stochastic* variational inference (SVI) methods (Hoffman et al., 2013). In our particular case, the bound $\mathcal{L}_{\mathcal{C}}$ is expressed as

$$\begin{aligned}\mathcal{L}_{\mathcal{C}} = \sum_{n=1}^{N_{\text{new}}} \mathbb{E}_{q(\mathbf{f}_n)}[\log p(y_n|\mathbf{f}_n)] - \text{KL}[q(\mathbf{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})||p(\mathbf{u}_{\text{new}}|\boldsymbol{\psi}_{\text{new}})] \\ + \text{KL}[q(\mathbf{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})||p(\mathbf{u}_{\text{new}}|\boldsymbol{\psi}_{\text{old}})] - \text{KL}[q(\mathbf{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})||\widetilde{q}(\mathbf{u}_{\text{new}}|\boldsymbol{\phi}_{\text{old}})].\end{aligned} \tag{3.36}$$

So far, under a factorized bound of this form, we are able to combine both continual learning with stochastic optimization, splitting our new incoming subset of data $\mathcal{D}_{\text{new}}$ in smaller *mini-batches* for faster training. Intuitively, it makes the $\mathcal{L}_{\mathcal{C}}$ bound applicable to a wider number of real-world problems, particularly those ones with an extremely asymmetric sequence of observations. That is, if the size of streaming batches is still large for training, we can apply SVI until the next incoming batch will be observed. Thus, the combinatio of SVI within the continual learning framework leads to a *best-of-both-worlds* strategy, since many times stochastic approximations can be also considered for streaming settings (Hensman et al., 2013a). In contrast, if the number of new observations goes to the opposite limit, i.e. a reduced number of samples per time-step t, then, the stochastic version of the $\mathcal{L}_{\mathcal{C}}$ bound in Eq. (3.36) can be avoided, leading to solutions closer to the work in Solin et al. (2018) and Bayesian filtering.

3.2.2 Generalization for Multi-task Models

Regarding the applicability of continual GP priors to high-dimensional output settings, we study how to adapt the previous results to sequences of multiple output data. Concretely in this section, we are interested in the generalisation of the continual GP scheme to accept extremely asymmetric cases. For instance, those ones for which, in addition to an unknown stream of observations, the order of appearance of the multi-output dimensions might be unknown as well. Several cases of both symmetric and asymmetric observation processes are depicted in Fig. 3.5.

We begin by considering parallel sequences with diferent size, formally denoted as *channels*, $\mathcal{D}_d$ with $d \in [1, \dots, D]$. From each dth channel, we sequentially observe batches of input-output data, such that $\mathcal{D}_d = \{\mathcal{Y}_d^{(1)}, \mathcal{Y}_d^{(2)}, \dots, \mathcal{Y}_d^{(t)}\}$ where $\mathcal{Y}_d^{(t)} = \{y_d(\boldsymbol{x}_n)_{n=1}^{N_d^t}\}$ and $\boldsymbol{x}_n \in \mathbb{R}^p$. Notice that here, time steps t are not necessarily aligned across different channels, and its size N_d^t may also vary. At this point, we initially consider the case for which each output value $y_d(\boldsymbol{x}_n)$ is continuous and Gaussian distributed. This asumption with be relaxed later on this section.

Having a multiple output problem of this type, we want to jointly model it using multi-output Gaussian processes (MOGP) as we did in Sec. 2.1.2 for heterogeneous likelihood problems. These models, typically generalise the flexible prediction system of GP approaches

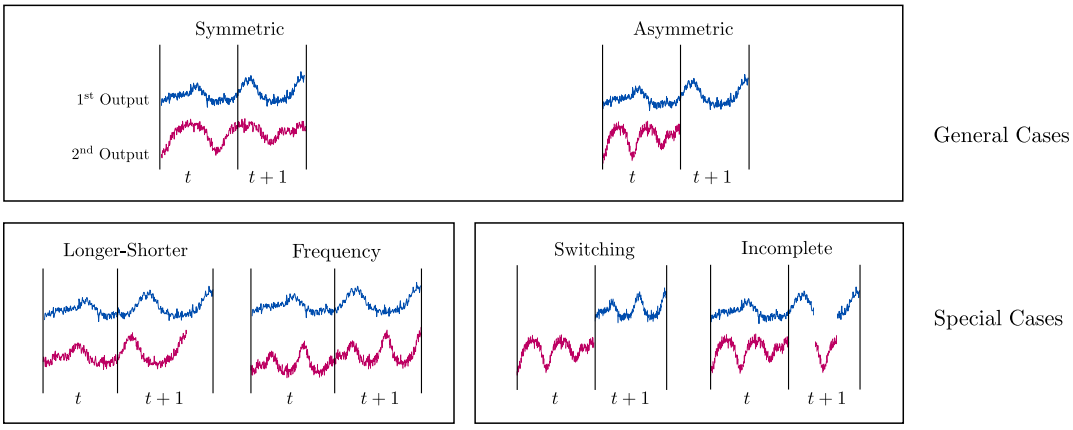


Figure 3.5: Illustration of the scenarios that two sequences of streaming input-output observations may belong to. Short-code for figures: (R = right, L = left). (UPPER ROW) General cases for the two output channels: symmetric (L) and asymmetric (R) sequential data. (LOWER ROW) Special forms of the upper cases: i) one channel is longer at time $t+1$ (L1), ii) channels have different frequency (L2), iii) switching missing channels (R1) and iv) both outputs sequences are in incomplete (R2).

to the vector-valued random field setup (Alvarez et al., 2012). In particular, it is demonstrated that by exploiting correlations among different streams of outputs, or channels, they are able to improve in the prediction for every d-th output. We aim to exploit this idea of correlated outputs in the multi-task sequential framework. However, little work has been done on extending MOGPs to the continual learning scenario. The most closely related works to this idea are Cheng et al. (2017) and Yang et al. (2018). Importantly, the model to be presented in this section is different from Cheng et al. (2017) because we allow for continual updates of the MOGP while they focus on adding structure to the kernel functions. The work by Yang et al. (2018) also derives tractable variational lower bounds based on the sparse approximation, but they do not handle non-Gaussian likelihoods and the learning method uses particle filtering with a *fixed* number of inducing points. In this section, we present a novel extension to perform continual learning given any MOGP model, independently of the likelihood distributions considered.

Sequential Multi-output Formulation

Having a multi-parameter GP prior as the one described in Ch. 2, we want to model the sequential observation process properly. Thus, suppose that we expect to observe a high-dimensional dataset $\mathcal{D} = \{\boldsymbol{x}_n, \boldsymbol{y}_n\}_{n=1}^{N}$ where we know *a priori* that output vectors $\boldsymbol{y}_n \in \mathbb{R}^{D\times 1}$ are composed by D features, such that $\boldsymbol{y}_n = [y_1(\boldsymbol{x}_n), y_2(\boldsymbol{x}_n), \ldots, y_D(\boldsymbol{x}_n)]^\top$ with $\boldsymbol{x}_n \in \mathbb{R}^p$ as in the single-output scenario. Again, we assume that the data $\mathcal{D}$ will be observed as a continual flow of smaller batches $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_t$ with irregular size and unknown arrival time.

We also assume that the pairs of input-output observations are aligned across channels, that is, the streaming setting is equivalent to the single-output case but considering the output vectors $\boldsymbol{y}_n$ instead of scalars for simplicity in the derivation. Importantly, the multi-output model presented here is also applicable to the case of asymmetric channels (see Fig. 3.5), as we will show later on this section.

The generative process of the multi-output samples is as follows. We assume that there

exist Q latent functions $\mathcal{U}$ that are linearly combined to produce D latent output functions $\mathcal{F}$ along time, using the previous LMC formulation of Sec. 2.2.1. However, in our MOGP prior, each one of the underlying $\mathcal{U}$ functions is *stationary* across batches $\mathcal{D}_t$ and their output variable $\boldsymbol{y}_n$ follows a probability distribution $p(\boldsymbol{y}_n|\boldsymbol{f}_n) = \prod_{d=1}^{D} p(y_d(\boldsymbol{x}_n)|f_d(\boldsymbol{x}_n))$.

Moreover, we also define the output vector $\boldsymbol{f}_n$ as $\boldsymbol{f}_n = [\boldsymbol{f}_1^\top, \boldsymbol{f}_2^\top, \ldots, \boldsymbol{f}_D^\top]^\top$, where $\boldsymbol{f}_n \in \mathbb{R}^{DN_t \times 1}$. We re-use the notation from the single-output case of Sec. 3.2.1 to indicate that our dataset is recursively partitioned, as $\mathcal{D} = \{\mathcal{D}_{\text{old}}, \mathcal{D}_{\text{new}}\}$, where $\mathcal{D}_{\text{new}} = \mathcal{D}_t$ at each time-step t and $\mathcal{D}_{\text{old}}$ ever increases. When training this MOGP model for exact inference, the problem is analogous to the continual GP case of Sec. 3.2.1. This is, we encounter a recurrent computational cost that now also includes D, the number of outputs, such that $\mathcal{O}(D^3N_1^3)$, $\mathcal{O}(D^3(N_1+N_2)^3), \ldots, \mathcal{O}(D^3N^3)$. Even if we avoid the use of non-Gaussian likelihoods for every output, where exact posterior inference is intractable, such computational cost is still unfeasible.

Sparse Approximations for Continual MOGP Models – We introduce inducing variables within variational inference methods for a reason of scalability. Sparse approximations have been already used in this context Alvarez and Lawrence (2009); Álvarez et al. (2010); Moreno-Muñoz et al. (2018). The subtle difference from the single-output case of Sec. 3.2.1 lies on the fact that pseudo-observations are not taken from the output functions $\mathcal{F}$ but from the latent ones $\mathcal{U}$ instead. Consequently, the extra layer that the multi-output GP adds for correlating latent functions is also used here for the sparse approximation. This induces a two-step conditioning on the model. For instance, the output function values are conditioned to the latent functions and at the same time, latent function vectors are conditioned to the subset of pseudo-observations in a *chained* form.

Under this setting, we define Q sets of M_q inducing variables, one per function $u_q(\cdot)$, such that $\boldsymbol{z} = \{\boldsymbol{z}_m\}_{m=1}^{M_q}$ and $\boldsymbol{z} \in \mathbb{R}^{M_q \times p}$. It is important to remark that these subsets are not restricted to take the same values $\boldsymbol{z}_m$ across dimensions and neither the sample size M_q must be equal for every q-th function. However, we do consider all M_q to be identical and equal to M in this work, mainly for simplicity in the notation. We also denote the vector $\boldsymbol{u}_q = [u_q(\boldsymbol{z}_1), u_q(\boldsymbol{z}_2), \ldots, u_q(\boldsymbol{z}_M)]^\top$ as the LF evaluations given the u_q process. Additionally, we set $\boldsymbol{u} = [\boldsymbol{u}_1^\top, \boldsymbol{u}_2^\top, \ldots, \boldsymbol{u}_Q^\top]^\top$ and $\boldsymbol{u} \in \mathbb{R}^{QM \times 1}$ for the whole set of functions $\mathcal{U}$. Notice that here, we have the sparse GP notation transformed for the multi-output problem.

Given D output functions $\mathcal{F}$ and the Q latent functions $\mathcal{U}$, we now aim to build our joint prior to be $p(\mathcal{F}, \mathcal{U}) = p(\mathcal{F}|\mathcal{U})p(\mathcal{U}|\boldsymbol{\psi})$, where once again, we use $\boldsymbol{\psi}$ to refer the subset of hyperparameters involved in the MOGP prior. Using the infinite-dimensional approach that we introduced in the previous single-output case, we now can factorize by conditioning on the finite number of inducing points $\boldsymbol{u}$. That is,

$$p(\mathcal{U}|\boldsymbol{\psi}) = p(\mathcal{U}_{\neq \boldsymbol{u}}|\boldsymbol{u}, \boldsymbol{\psi})p(\boldsymbol{u}|\boldsymbol{\psi}), \tag{3.37}$$

where $\mathcal{U}_{\neq \boldsymbol{u}}$ refers to all latent function values $\mathcal{U}$ not including that ones explored in $\boldsymbol{u}$, that is, $\mathcal{U} = \mathcal{U}_{\neq \boldsymbol{u}} \cup \boldsymbol{u}$. The prior distribution over $\boldsymbol{u}$ also factorizes across latent functions, as $p(\boldsymbol{u}|\boldsymbol{\psi}) = \prod_{q=1}^{Q} p(\boldsymbol{u}_q|\boldsymbol{\psi})$ with $\boldsymbol{u}_q \sim \mathcal{N}(\boldsymbol{0}, \mathbf{K}_q)$. Here, $\mathbf{K}_q \in \mathbb{R}^{M \times M}$ corresponds to $k_q(\boldsymbol{z}_i, \boldsymbol{z}_j)$ with entries $\boldsymbol{z}_i, \boldsymbol{z}_j \in \boldsymbol{z}$. The dimension of $\mathbf{K}_q$ varies within the number of inducing points evaluations, determining the model's maximum complexity. This last point plays an important role when the input domain is incremented within the appearance of newer observations.

Continual Multi-output Inference

Our primary goal is to obtain the posterior distribution $p(\boldsymbol{f}, \boldsymbol{u}|\mathcal{D})$, that we know *a priori* is intractable under the presence of inducing points and potential non-Gaussian likelihoods.

If we consider the variational approach, once again, as in the approach of Titsias (2009b), where we can approximate our posterior with an auxiliary Gaussian distribution $q(\cdot,\cdot)$, we may consider the following factorisation as in Álvarez et al. (2010),

$$p(\boldsymbol{f},\boldsymbol{u}|\mathcal{D}) \approx q(\boldsymbol{f},\boldsymbol{u}) = p(\boldsymbol{f}|\boldsymbol{u})q(\boldsymbol{u}) = \prod_{d=1}^{D} p(\boldsymbol{f}_d|\boldsymbol{u}) \prod_{q=1}^{Q} q(\boldsymbol{u}_q), \tag{3.38}$$

where we have used a product of Q Gaussian densities, one per qth latent process, with $q(\boldsymbol{u}_q) = \mathcal{N}(\boldsymbol{u}_q|\boldsymbol{\mu}_{\boldsymbol{u}_q}, \mathbf{S}_{\boldsymbol{u}_q})$ and where the conditional probabilities $p(\boldsymbol{f}_d|\boldsymbol{u})$ are given by

$$p(\mathbf{f}_d|\mathbf{u}) = \mathcal{N}\Big(\boldsymbol{f}_d|\mathbf{K}_{\boldsymbol{f}_d\boldsymbol{u}}\mathbf{K}_{\boldsymbol{u}\boldsymbol{u}}^{-1}\boldsymbol{u}, \mathbf{K}_{\boldsymbol{f}_d\boldsymbol{f}_d} - \mathbf{K}_{\boldsymbol{f}_d\boldsymbol{u}}\mathbf{K}_{\boldsymbol{u}\boldsymbol{u}}^{-1}\mathbf{K}_{\boldsymbol{f}_d\boldsymbol{u}}^{\top}\Big). \tag{3.39}$$

The cross-covariance matrices $\mathbf{K}_{\boldsymbol{f}_d\boldsymbol{u}} \in \mathbb{R}^{N\times QM}$ are obtained by direct evaluation of the correlation between $\boldsymbol{f}_d(\boldsymbol{x})$ and $u_q(\boldsymbol{z})$. We also denote $\mathbf{K}_{\boldsymbol{u}\boldsymbol{u}} \in \mathbb{R}^{QM\times QM}$ as the block-diagonal matrix formed by the $\mathbf{K}_q$ terms.

Avoiding Revisiting Multiple Likelihoods

When using variational inference methods, we fit the auxiliary distribution $q(\boldsymbol{u}_q)$ by maximising a lower-bound $\mathcal{L}$ of the log-marginal likelihood $p(\mathcal{D})$. In the MOGP literature, this marginal is also written as $\log p(\boldsymbol{y})$ and in our case, we express it also as $\log p(\boldsymbol{y}_{\text{new}}, \boldsymbol{y}_{\text{old}})$. Given the previously defined sparse MOGP model, this probability distribution can be decomposed as a double integral operator,

$$\log p(\boldsymbol{y}_{\text{new}}, \boldsymbol{y}_{\text{old}}) = \log \iint p(\boldsymbol{y}_{\text{new}}, \boldsymbol{y}_{\text{old}}|\mathcal{F}) p(\mathcal{F},\mathcal{U}) d\mathcal{F} d\mathcal{U}, \tag{3.40}$$

where we now consider the finite set of output values $\boldsymbol{y}_{\text{old}}$ and $\boldsymbol{y}_{\text{new}}$ to be conditioned on the infinite-dimensional domain of the set of functions $\mathcal{F}$, as is done in Bui et al. (2017b) and Moreno-Muñoz et al. (2020a), but in the multi-output counterpart. Due to this assumption, we have a double expectation over both $\mathcal{F}$ and $\mathcal{U}$, where we can also apply conditional independence (CI) in the main likelihood term of Eq. (3.40). This CI leads us to obtain $p(\boldsymbol{y}_{\text{new}}, \boldsymbol{y}_{\text{old}}|\mathcal{F}) = p(\boldsymbol{y}_{\text{new}}|\mathcal{F})p(\boldsymbol{y}_{\text{old}}|\mathcal{F})$. For simplicity, we will denote both terms as the *new* and *old* likelihoods respectively.

As it was previously mentioned, when dealing with variational inference, any lower bound $\mathcal{L}$ over Eq. (3.40) requires to sequentially evaluate expectatons given former log-likelihood terms $\log p(\boldsymbol{y}_{\text{old}}|\boldsymbol{f})$. However, under the assumption of a multi-output GP model, the recurrent evaluation of expectations even worsens. In particular, due to the factorization of LPFs, it is necessary to compute, at least, D integrals over the dimensions of the *old* data vectors $\boldsymbol{y}_{\text{old}}$. Notice that each d-th dimension might be characterized by a different likelihood function, that we also aim to estimate. Fortunately, the solution in Bui et al. (2017a) still yields in the multiple channel setting. As in Sec. 3.2.1, we can approximate all probabilities $p(\boldsymbol{y}_{\text{old}}|\boldsymbol{f})$ by means of the Bayes rule. We have that as long as,

$$q(\mathcal{F},\mathcal{U}) \approx p(\mathcal{F},\mathcal{U}|\boldsymbol{y}_{\text{old}}, \boldsymbol{x}_{\text{old}}) \propto p(\mathcal{F},\mathcal{U})p(\boldsymbol{y}_{\text{old}}|\mathcal{F}), \tag{3.41}$$

we can invert the Bayes rule to obtain an unnormalized estimate of the past likelihood term $p(\boldsymbol{y}_{\text{old}}|\mathcal{F})$ as

$$p(\boldsymbol{y}_{\text{old}}|\mathcal{F}) \approx \frac{q(\mathcal{F},\mathcal{U})}{p(\mathcal{F},\mathcal{U})}. \tag{3.42}$$

Importantly, the two given distributions that intervene in the quotient of Eq. (3.42) factorise according to

$$q(\mathcal{F},\mathcal{U}) = p(\mathcal{F}|\mathcal{U})p(\mathcal{U}_{\neq \boldsymbol{u}}|\boldsymbol{u},\boldsymbol{\psi}_{\text{old}}) \prod_{q=1}^{Q} q(\boldsymbol{u}_q), \tag{3.43}$$

$$p(\mathcal{F},\mathcal{U}) = p(\mathcal{F}|\mathcal{U})p(\mathcal{U}_{\neq \boldsymbol{u}}|\boldsymbol{u},\boldsymbol{\psi}_{\text{old}}) \prod_{q=1}^{Q} p(\boldsymbol{u}_q|\boldsymbol{\psi}_{\text{old}}), \tag{3.44}$$

where both variational posterior distributions $q(\cdot)$ and priors $p(\cdot)$ are evaluated over the inducing points given by the respective Q latent functions. This fact makes easier to obtain separated KL divergence terms in the future continual lower bound for multi-task problems.

Additionally, if we introduce the aforementioned expression in Eq. (3.42), as a sequential estimator of the multiple *old* likelihood terms, we can reformulate Eq. (3.40) to be

$$\log p(\boldsymbol{y}_{\text{new}}, \boldsymbol{y}_{\text{old}}) \approx \log \iint \frac{p(\boldsymbol{y}_{\text{new}}|\mathcal{F})p(\mathcal{F},\mathcal{U})q(\mathcal{F},\mathcal{U})}{p(\mathcal{F},\mathcal{U})} d\mathcal{F}d\mathcal{U}, \tag{3.45}$$

where both prior densities $p(\mathcal{F},\mathcal{U})$ in the quotient differs between them due to different subsets of hyperparameters, i.e. $\boldsymbol{\psi}_{\text{old}}$ vs. $\boldsymbol{\psi}_{\text{new}}$. Having an approximated log-marginal distribution of this form, we can build the lower bound via Jensen's inequality, that is,

$$\mathcal{L} = \iint q(\mathcal{F},\mathcal{U}|\boldsymbol{\phi}_{\text{new}}) \log \frac{p(\boldsymbol{y}_{\text{new}}|\mathcal{F})p(\mathcal{F},\mathcal{U})q(\mathcal{F},\mathcal{U})}{q(\mathcal{F},\mathcal{U}|\boldsymbol{\phi}_{\text{new}})p(\mathcal{F},\mathcal{U})} d\mathcal{F}d\mathcal{U}. \tag{3.46}$$

In this MOGP setting, there is also a problem related to the use of past *explicit* distributions in the bound in Eq. (3.46) as we saw in Sec. 3.2.1. This issue remains as we have to propagate past subsets of instances $\boldsymbol{u}_{\text{old}}$ forward, for each latent function, in order to approximate the likelihood probabilities. To avoid it, we also adapted the continual GP model within the predictive expressions presented in this section.

Algorithm 2 — MULTI-CHANNEL CONTINUAL GP LEARNING

Initialize $\boldsymbol{\phi}_{\text{new}}^{(0)}$ and $\boldsymbol{\psi}_{\text{new}}^{(0)}$ randomly.
input: Observe $\mathcal{D}_{\text{new}}^{(0)}$
Maximise $\mathcal{L} \leq \log p(\mathcal{D}_{\text{new}}^{(0)})$ w.r.t. $\{\boldsymbol{\phi}_{\text{new}}^{(0)}, \boldsymbol{\psi}_{\text{new}}^{(0)}\}$. // standard variational inference
for $t \in 1, \ldots, T$ **do**
  Update $\{\boldsymbol{\phi}_{\text{old}}^{(t)}, \boldsymbol{\psi}_{\text{old}}^{(t)}\} \leftarrow \{\boldsymbol{\phi}_{\text{new}}^{(t-1)}, \boldsymbol{\psi}_{\text{new}}^{(t-1)}\}$ // fitted parameters are repurposed
  for $q \in 1, \ldots, Q$ **do**
    input: Observe $\mathcal{D}_{\text{new}}^{(t)}$
    Choose initial $\boldsymbol{Z}_{\text{new}}$ // initialization of inducing points
    Compute continual GP priors $\widetilde{q}(\cdot|\boldsymbol{\phi}_{\text{old}}^{(t)})$ // conditional prior reconstruction
  end for
  Maximise $\mathcal{L}_{\mathcal{C}}$ w.r.t. $\{\boldsymbol{\phi}_{\text{new}}^{(t)}, \boldsymbol{\psi}_{\text{new}}^{(t)}\}$. // continual variational inference
end for

Consider an additional set of inducing inputs $\mathcal{Z}_*$, that we will use as instances of the latent functions $\mathcal{U}$ in the MOGP prior. Thus, assuming that $p(\boldsymbol{u}|\mathcal{D}) \approx q(\boldsymbol{u})$, the predictive distribution $p(\mathcal{U}_*|\mathcal{D})$ can be approximated as $\int p(\mathcal{U}_*|\boldsymbol{u})q(\boldsymbol{u})d\boldsymbol{u}$. Here, we use the variable $\mathcal{U}_*$ to denote the LF instances taken on $\mathcal{Z}_*$. While $q(\boldsymbol{u})$ factorises accross the Q latent function

vectors $\boldsymbol{u}_q$, the conditional MOGP prior $p(\mathcal{U}_*|\boldsymbol{u})$ is analogous to the one presented in Eq. (3.27). This means that we can apply the same predictive mechanism to build continual MOGP priors, that now works in the subset of LFs before the mixing.

As a consequence, for each q-th latent process u_q, we obtain a continual GP prior of the form $\widetilde{q}(u_*|\boldsymbol{\phi}_{\text{old}}) \approx \int p(u_*|\boldsymbol{u}_q)q(\boldsymbol{u}_q|\boldsymbol{\phi}_{\text{old}})d\boldsymbol{u}_q$. Additionally, due to the latent processes are independent, having their own covariance functions, the continual update is also independent. In particular, we assume the existence of Q parallel continual processes of the form

$$u_{q,*} \sim \mathcal{GP}(k_{*\boldsymbol{u}_q}\mathbf{K}_{\boldsymbol{u}_q\boldsymbol{u}_q}^{-1}\boldsymbol{\mu}_{q,\text{old}}, k_{**} + k_{*\boldsymbol{u}_q}\mathbf{K}_{\boldsymbol{u}_q\boldsymbol{u}_q}^{-1}(\mathbf{S}_{q,\text{old}} - \mathbf{K}_{\boldsymbol{u}_q\boldsymbol{u}_q})\mathbf{K}_{\boldsymbol{u}_q\boldsymbol{u}_q}^{-1}k_{*\boldsymbol{u}_q}^{\top}), \tag{3.47}$$

where $k_{*\mathbf{u}_q} = [k_q(\cdot, \boldsymbol{z}_1), \cdots, k_q(\cdot, \boldsymbol{z}_{M_q})]^{\top}$ refers to the values taken by the corresponding kernel. The development of the multi-output continual bound is now feasible. If we use the predictive GP equation to factorize $q(\mathcal{F},\mathcal{U}) = p(\mathcal{F}|\mathcal{U})p(\mathcal{U}_{\neq\mathbf{u}}|u_{q,*}, \boldsymbol{\psi}_{\text{old}})\prod_{q=1}^{Q} q(u_{q,*})$, then we can introduce $u_{q,*} = \mathbf{u}_{q,\text{new}}$ recursively. This leads to

$$\mathcal{L} = \iint q(\mathcal{F},\mathcal{U}|\boldsymbol{\phi}_{\text{new}}) \log \frac{p(\boldsymbol{y}_{\text{new}}|\mathcal{F})p(\mathbf{u}_{\text{new}}|\boldsymbol{\psi}_{\text{new}})}{q(\mathbf{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})} d\mathcal{F}d\mathcal{U} + \iint q(\mathcal{F},\mathcal{U}) \log \frac{\widetilde{q}(\mathbf{u}_{\text{new}}|\boldsymbol{\phi}_{\text{old}})}{p(\mathbf{u}_{\text{new}}|\boldsymbol{\psi}_{\text{old}})} d\mathcal{F}d\mathcal{U},$$

that we also rewrite in a more recognizable form, as

$$\begin{aligned}\mathcal{L} =& \sum_{d=1}^{D} \mathbb{E}_{q(\boldsymbol{f}_{d,\text{new}})}\left[\log p(\boldsymbol{y}_{d,\text{new}}|\boldsymbol{f}_{d,\text{new}})\right] - \sum_{q=1}^{Q} \text{KL}\left[q(\boldsymbol{u}_{q,\text{new}}|\boldsymbol{\phi}_{\text{new}})||p(\boldsymbol{u}_{q,\text{new}}|\boldsymbol{\psi}_{\text{new}})\right] \\ &+ \sum_{q=1}^{Q} \text{KL}\left[q_{\text{new}}(\boldsymbol{u}_{q,\text{new}}|\boldsymbol{\phi}_{\text{new}})||p(\boldsymbol{u}_{q,\text{new}}|\boldsymbol{\psi}_{\text{old}})\right] - \sum_{q=1}^{Q} \text{KL}\left[q(\boldsymbol{u}_{q,\text{new}}|\boldsymbol{\phi}_{\text{new}})||\widetilde{q}(\boldsymbol{u}_{q,\text{new}}|\boldsymbol{\phi}_{\text{old}})\right],\end{aligned} \tag{3.48}$$

where $q(\boldsymbol{f}_{d,\text{new}}) = \mathbb{E}_{q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})}[p(\mathbf{f}_{d,\text{new}}|\boldsymbol{u}_{\text{new}})]$ is the approximate marginal posterior for every $\boldsymbol{f}_{d,\text{new}} = f_d(\boldsymbol{x}_{\text{new}})$ that can be obtained analytically via

$$\begin{aligned}q(\mathbf{f}_{d,\text{new}}) = \mathcal{N}(\boldsymbol{f}_{d,\text{new}}|&\mathbf{K}_{\boldsymbol{f}_{d,\text{new}}\boldsymbol{u}_{\text{new}}}\mathbf{K}_{\boldsymbol{u}_{\text{new}}\boldsymbol{u}_{\text{new}}}^{-1}\boldsymbol{\mu}_{\boldsymbol{u}_{\text{new}}}, \mathbf{K}_{\boldsymbol{f}_{d,\text{new}}\boldsymbol{f}_{d,\text{new}}} \\ &+ \mathbf{K}_{\boldsymbol{f}_{d,\text{new}}\boldsymbol{u}_{\text{new}}}\mathbf{K}_{\boldsymbol{u}_{\text{new}}\boldsymbol{u}_{\text{new}}}^{-1}(\mathbf{S}_{\boldsymbol{u}_{\text{new}}} - \mathbf{K}_{\boldsymbol{u}_{\text{new}}\boldsymbol{u}_{\text{new}}})\mathbf{K}_{\boldsymbol{u}_{\text{new}}\boldsymbol{u}_{\text{new}}}^{-1}\mathbf{K}_{\boldsymbol{f}_{d,\text{new}}\boldsymbol{u}_{\text{new}}}^{\top}),\end{aligned} \tag{3.49}$$

where $\boldsymbol{\mu}_{\boldsymbol{u}_{\text{new}}} = [\boldsymbol{\mu}_{\boldsymbol{u}_{1,\text{new}}}^{\top}, \cdots, \boldsymbol{\mu}_{\boldsymbol{u}_{Q,\text{new}}}^{\top}]$ and $\mathbf{S}_{\boldsymbol{u}_{\text{new}}}$ is a block matrix whose elements are given by $\mathbf{K}_{\boldsymbol{u}_{q,\text{new}}}$. The intepretability of Eq. (3.48) is of particular interest in our work. In the standard GP case, both expectations and divergence terms refer to the same layer of computation. This is, integrals calculate the uncertainty on the output function instances of $f(\cdot)$, the one which parameterises the likelihood densities. However, in the MOGP setting, it is different. Particularly, the expectation term in Eq. (3.48) is focused in the observations and the parameter output functions. On the other hand, the three KL divergences affect exclusively to the layer of latent processes $\mathcal{U}$. This point makes the continual method applicable to *asymmetric* scenarios or where, for instance, one of the output channels might be unobserved at a time-step t. This will be analyzed in the experiments.

Moreover, in Alg. 2, we present all necessary computations for the continual learning of the model. The algorithm is analogous to the one presented in Alg. 1, but in this case, we require Q extra iterations over the LFs.

3.3 Evaluation of Continual and Distributed Inference

We now examine the empirical performance of the GP models for *continual* and *distributed* inference. In the first case, we evaluate the continual GP model in both single-output and

multi-output scenarios. Results of evaluations are presented in two forms, one more visual via the representation of the GP test prediction in the input area of interest, and another based on predictive error metrics, e.g. negative log-predictive density NLPD. Further details about the general implementation of experiments, initial setting of hyperparameters or access to data, are provided in the appendix of this thesis. The implementation of both models is publicly available in the repository https://github.com/pmorenoz/.

3.3.1 Recyclable GP Simulations

In this section, we present the simulations for analising the performance of our *distributed* framework for multiple recyclable GP models and data access settings. To illustrate its usability, and also in the context of this thesis, we show results in three different learning scenarios: i) regression, ii) classification and iii) heterogeneous data. Performance metrics are given in terms of the negative log-predictive density (NLPD), root mean square error (RMSE) and mean-absolute error (MAE). The experiments were programmed in Pytorch[1], which allows to learn the GP ensembles in an automatic manner as well as the baseline methods used. Importantly, we remark that data is never revisited and its presence in the ensemble plots is just for clarity in the presentation of results.

Concatenation Test with Toy Data

For the first experiment, whose results are illustrated in Fig. 3.6. We generated $K = 5$ subsets of observations in the input-space range $\boldsymbol{x} \in [0.0, 5.5]$. Each subset contains $N_k = 500$ uniform samples of $\boldsymbol{x}_k$. These were later evaluated as instances $f(\boldsymbol{x})_{\text{bias}}$. The expression of the true function is

$$f(\boldsymbol{x})_{\text{bias}} = f(\boldsymbol{x}) + 3\boldsymbol{x} - \frac{15}{2}, \tag{3.50}$$

where the expression for the used *unbiased* mapping is

$$f(\boldsymbol{x}) = \frac{9}{2}\cos\left(2\pi\boldsymbol{x} + \frac{3\pi}{2}\right) - 3\sin\left(\frac{43\pi}{10}\boldsymbol{x} + \frac{3\pi}{10}\right). \tag{3.51}$$

Having the local values of the true underlying function $\boldsymbol{f}_k = f(\boldsymbol{x}_k)$, we generated the true output targets using additive Gaussian noise, such that $\boldsymbol{y}_k = \boldsymbol{f}_k + \epsilon_k$, where $\epsilon_k \sim \mathcal{N}(0, 2)$. For each local task, the number of inducing-inputs $\boldsymbol{Z}_k$ was $M_k = 15$ and their initialization was equally spaced in the corresponding local input regions. For the global GP prediction, we used $M = 35$ inducing-inputs $\boldsymbol{Z}_*$, initialized in the same manner. In Fig. 3.6, we show three of five tasks united in a new GP model. Tasks are fitted independently with $N_k = 500$ synthetic data points per distributed GP. Notice that the variational ensemble tends to match the uncertainty of the local approximations.

Distributed Gaussian Process Regression

We provide error metrics for the recyclable GP framework compared with the state-of-the-art models in Tab. 3. The training dataset is also synthetic in this experiment and generated using the expressions above. For the case with 10K observations, we used $K = 50$ tasks with $N_k = 200$ data-points and $M_k = 3$ inducing variables in the sparse GP. The scenario for $N = 100$K is similar but divided into $K = 250$ tasks with $N_k = 400$. Our method obtains better results that the exact distributed solutions due to the ensemble bound searches the

[1] PYTHON framework with auto-differentiation available at https://pytorch.org/.

Figure 3.6: Recyclable GPs with toy data.

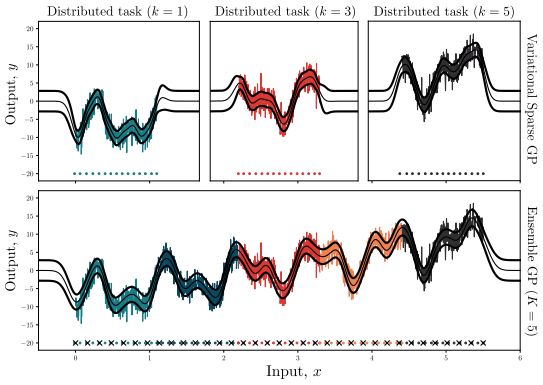


Table 2: Performance metrics for distributed GPs with the solar physics dataset. (std. $\times 10^2$)

MODEL	NLPD	RMSE
BCM	–	17.25
PoE	1.51 ± 0.01	1.08
GPoE	1.51 ± 0.07	1.08
RBCM	1.53 ± 0.01	1.11
This work	1.68 ± 0.14	1.17 ± 0.12

DATA SIZE →	10K			100K			1M		
MODEL	NLPD	RMSE	MAE	NLPD	RMSE	MAE	NLPD	RMSE	MAE
BCM	2.99 ± 0.94	11.94 ± 18.89	2.05 ± 1.31	3.51 ± 0.73	2.33 ± 0.96	1.34 ± 1.03	NA	9.56 ± 14.87	1.19 ± 0.86
PoE	2.79 ± 0.16	2.32 ± 0.22	1.86 ± 0.22	2.82 ± 0.67	2.19 ± 0.91	1.71 ± 0.84	2.91 ± 0.63	1.98 ± 0.61	1.32 ± 0.05
GPoE	2.79 ± 0.56	2.43 ± 0.52	1.96 ± 0.48	2.73 ± 0.72	2.19 ± 0.91	1.71 ± 0.84	2.72 ± 0.52	1.98 ± 0.61	1.32 ± 0.05
RBCM	2.96 ± 0.51	2.49 ± 0.51	2.02 ± 0.46	3.03 ± 0.86	2.51 ± 1.12	1.99 ± 1.04	2.56 ± 0.06	1.82 ± 0.02	1.37 ± 0.03
This work	2.71 ± 0.11	1.56 ± 0.04	0.97 ± 0.05	2.89 ± 0.07	1.73 ± 0.01	1.23 ± 0.02	2.87 ± 0.09	1.87 ± 0.07	1.34 ± 0.09

Acronyms: BCM (Tresp, 2000), PoE (Ng and Deisenroth, 2014), GPoE (Cao and Fleet, 2014) and RBCM (Deisenroth and Ng, 2015).

Table 3: Comparative error metrics for distributed GP models. Best values are underlined.

average solution among all recyclable GPs. The baseline methods are based on a combination of solutions, if one is bad-fitted, it has a direct effect on the predictive performance.

We also tested the data with the inference setup of Gal et al. (2014), obtaining an NLPD of 2.58 ± 0.11 with 250 nodes for 100K observations. It is better than our approach and the baseline methods, but without any GP reconstruction, only distributes the computational cost of products of matrices or/and their inversions.

Recyclable Ensembles

For a large synthetic dataset, for instance $N{=}10^6$, we tested the recyclable GPs with $K = 5 \cdot 10^3$ tasks as shown in Tab. 3. However, if we ensemble large amounts of local GPs, e.g. $K{\gg}10^3$, it is problematic for most of baseline methods, due to partitions must be revisited for building predictions and if one-of-many GP fails, performance decreases. Thus, we repeated the experiment in a *pyramidal* way. This is, building ensembles of recyclable ensembles, inpired in the approach of Deisenroth and Ng (2015). Our method obtains $\{\text{NLPD} = 4.15, \text{RMSE} = 2.71, \text{MAE} = 2.27\}$. The results in Tab. 3 indicate that our model is more robust under the *concatenation* of approximations rather than overlapping them in the input space. The *pyramidal* experiment was formed by two *layers*, that is, we joined ensembles twice as shown in Fig. 3.8.

Solar Physics Dataset

We also tested the framework on solar data (available at https://solarscience.msfc.nasa.gov/), which consists of more than $N{=}10^3$ monthly average estimates of the *sunspot*

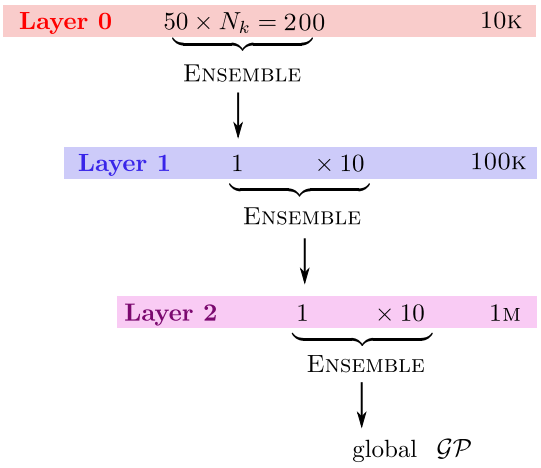


Figure 3.8: Graphical depiction of the *pyramidal* structure for ensembles of GP ensembles.

counting numbers from 1700 to 1995. We applied the mapping $\log(1 + y_n)$ to the output targets for performing Gaussian regression. Error metrics are provided in Tab. 2, where std. values were sufficiently small, so we do not include them. The performance with K=50 tasks is close to the baseline solutions, but without storing all distributed subsets of data. The number of global inducing-inputs used for the ensemble was M=90, whist we used M_k=6 for each distributed approximation.

Pixel-wise MNIST Classification

For this experiment, we took images of *ones* and *zeros* from the well-known MNIST dataset, inspired in the MNIST experiments in Van der Wilk et al. (2017). To simulate a pixel-wise unsupervised classification problem, true labels of images were ignored. Instead, we threshold the pixels of images to be greater or smaller than 0.5, and labeled as $y_n = 0$ or $y_n = 1$ afterwards. This is, we turned the grey-scaled values to a binary coding. Then, to simulate a pixel-wise scenario, we used each pixel as an input-output datum whose input $\boldsymbol{x}_n$ contains the two coordinates (x_1 and x_2 axes). Plots in Fig. 3.9 illustrate that a predictive ensemble can be built from smaller pieces of GP models, four corners in the case of the number *zero* and two for the number *one*.

Compositional Number Prediction

As an illustration of potential applications of the recyclable GP approach, we build a number *eight* predictor using exclusively two subsets of the approximations learned in the previous experiment with the image of the number *zero*. We used the $K = 4$ distributed tasks of the experiment with number *zero* and replicated the objects $\mathcal{E}_k$ twice. Then, the final input list to the ensemble was $\{\mathcal{E}_1, \ldots, \mathcal{E}_4, \mathcal{E}_5, \ldots, \mathcal{E}_8\}$. The set of partitions $\{\mathcal{E}_5, \ldots, \mathcal{E}_8\}$ was identical to the previous ones but we *shifted* their corresponding inducing-inputs $\boldsymbol{Z}_k$ by adding 1.2 in the vertical axis. This is, with smaller distributed tasks of two number *zeros*, we generated an ensemble of a number *eight*. We remark that this experiment is purely illustrative to show the potential uses of the framework in compositional learning applications.

Banana Dataset

The *banana* experiment is perhaps one of the most used datasets for testing GP classification models. We followed a similar strategy as the one used in the MNIST experiment. After removing the 33% of samples for testing, we partitioned the input area in four quadrants, i.e. as shown in Fig. 3.9 (C). For each partition, we set a grid of $M_k = 9$ inducing-inputs and later, the maximum complexity of the global sparse model was set to $M = 25$. The baseline GP classification method also used $M = 25$ pseudo-observations and obtained a NLPD value of $7.29 \pm 7.85 \cdot 10^{-4}$ after ten trials with different initializations. Our method obtained a test NLPD of 7.21 ± 0.04. This difference is understandable as the recyclable GP framework used a total amount of 4×16 inducing-inputs, that captured more uncertainty than the 16 of the baseline classifier.

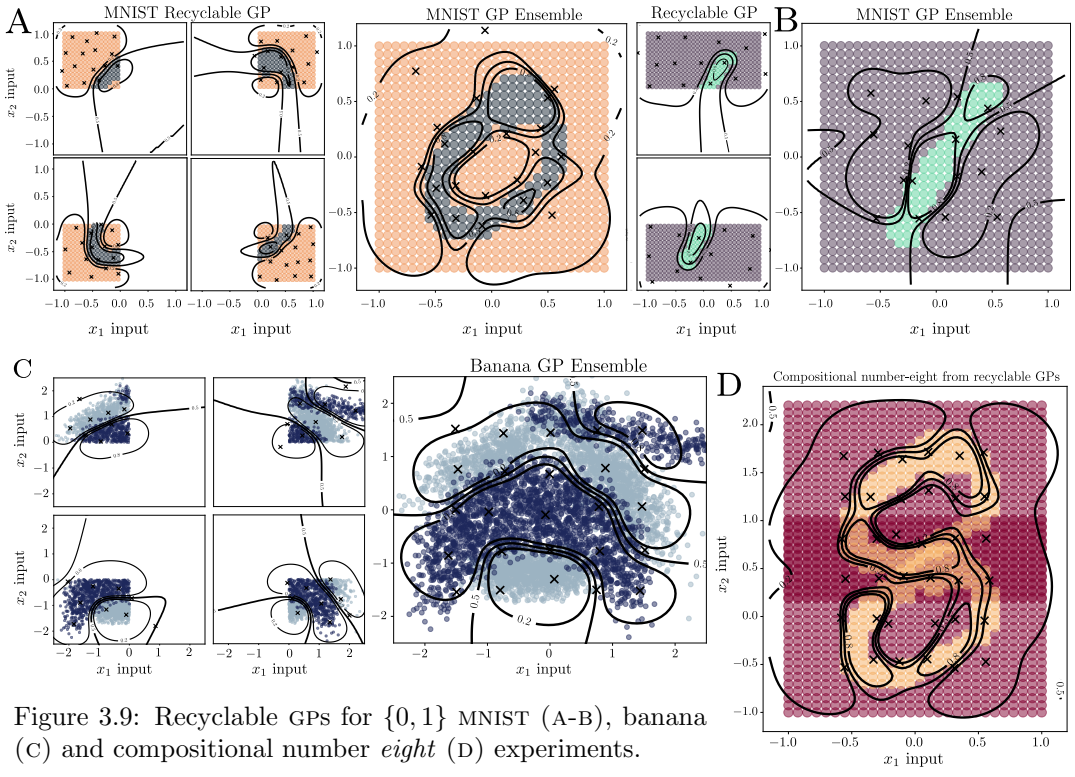


Figure 3.9: Recyclable GPs for $\{0, 1\}$ MNIST (A-B), banana (C) and compositional number *eight* (D) experiments.

3.3.2 Continual Multi-task GP Simulations

Our experiments in this section aim to examine the empirical performance of the continual GP approach. The results are focused in three main topics that demonstrate the utility and performance of the model, in particular over synthetic and real-world datasets. The three topics are: i) performance of the continual single-output GP model given streaming observations, ii) resistance to the error propagation when refitting variational approximations, including the appearance of new tasks, non-Gaussian observations and heterogeneous multi-output settings, iii) the applicability to real-world problems with multivariate *online* data, potentially configured as asymmetric channels. The experiments are also organized into several sub-topics, related to regression, classification, multi-view settings and last, heterogeneous likelihood problems.

For all experiments, we used a modified version of the PYTHON code released within Moreno-Muñoz et al. (2018) that presents similar features of scalability and adaptibility to multi-output and non-Gaussian data. For the optimization process w.r.t. continual lower bounds, we make use of the LBFGS-B algorithm (Zhu et al., 1997). If the stochastic counterpart was necessary, we considered ADADELTA instead (Zeiler, 2012), which is included in the *climin* library (Bayer et al., 2016). Further details about the general setting of experiments are included in the appendix of this thesis. Moreover, we provide a public repository (https://github.com/pmorenoz/ContinualGP/) where all scripts for simulations are provided for a reason of reproducibility.

Continual GP Regression

In this subset of experiments, we evaluate the performance of the continual GP model in single-output scenarios where streaming data is real-valued and assumed Gaussian distributed. Our goal is to perform sequential non-linear regression. We first setup a synthetic experiment with two different configurations in the way of appearance of incoming samples. These are denoted as i) streaming and ii) overlapping data. In the first case, we have a sequence of $t = 10$ non-overlapping partitions that are exclusively delivered to the learning system. Each partition in the collection avoids revisiting the previously explored input area. Second, we relax this assumption to consider partially overlapping tasks where certain regions in the input space a revisited by the new observations. Importantly, we always model a single-output latent function to link the likelihood parameters $\boldsymbol{\theta}_n$. This is, we avoid solutions similar to chained GPs (Saul et al., 2016), where several functions are used for parameterization for a reason of flexibility. However, this approach could be also considered to the current experiment with continual GPs.

i). STREAMING – The streaming data experiment consists of $t = 10$ partitions that are observed in a sequential manner. In this case, we consider that batches have an approximately equal size, so the scenario is not *irregular*. We setup the initial number of inducing inputs to be $M = 3$. This number sets the maximum complexity of the model. It is increased following the rule $M(t) = 3t$. The synthetic dataset has $N = 3000$ input-output samples, where 30% of them are used for testing errors.

In Fig. 3.10, we show three captions of the iterative learning process. Concretely, the initial step at $t = 1$, the intermediate one at $t = 5$ and the final step at $t = 10$. We remark that the posterior predictive computation of curves does not employ any past parameters, only the ones learned in the most recent iteration. The last trained model, whichs avoids to revisit data, is the one that predicts all along the input space explored so far. Additionally, in Tab. 3.3 we include the NLPD error values obtained at each iteration and sequential data collection. All posterior predictive densities are estimated via Monte-Carlo methods. The performance of the continual model is evaluated in three different ways. First, we evaluate test predictions at the new-observed input area. Second, we look to the *decay* of the predictive precision in the tth past seen input partitions. Finally, we evaluate which is the GP prediction quality along the whole input space in the experiment.

For instance, in the case of the column at $t' = 1$ in Tab. 3.3, the GP is trained at time steps $t = 1, 2, 3 \ldots$, until $t = 10$ (rows). Then, the continual process is evaluated on the test input data first seen at $t = 1$. This is, we want to evaluate how much precision is lost over the testing data stored at $t = 1$, t steps forward in the sequential learning. Therefore, the columns at $t' = 4$ and $t' = 8$ have a similar interpretation. The GP model first trained at $t = 5$ and then updated at $t = 6, 7, 8, \ldots$, until $t = 10$ are tested on the data first observed at $t = 4$. One can see how the *red* error metrics remain approximately static around an average NLPD

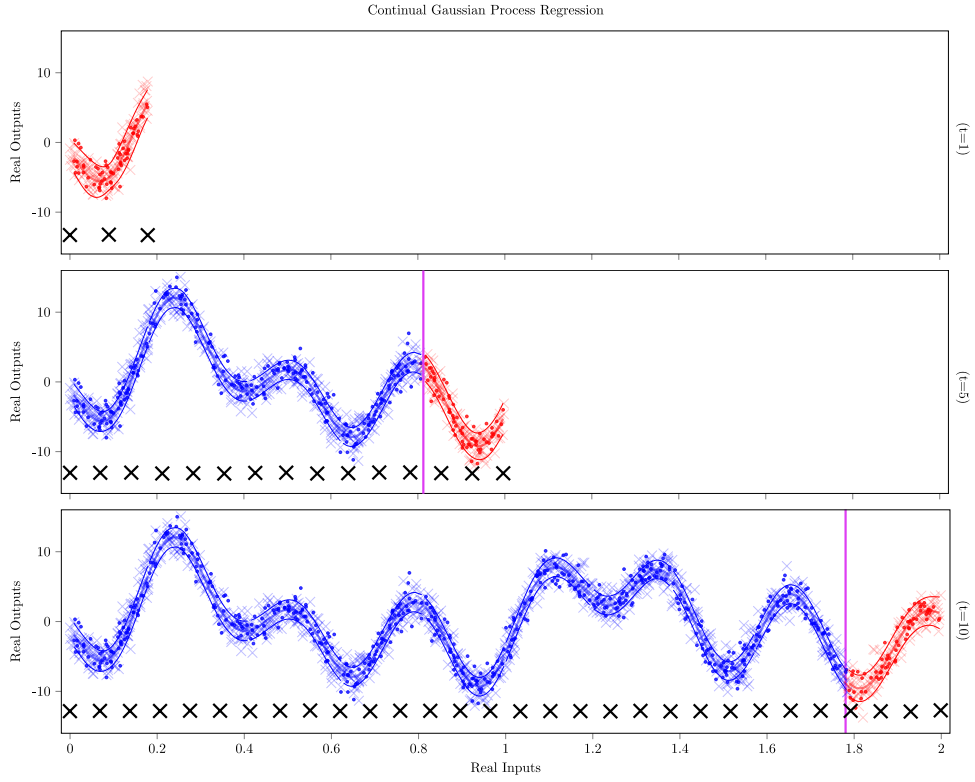


Figure 3.10: Results from continual GP regression applied to toy streaming data. Sequential batches correspond to non-overlapping partitions. The sequence consists of $t = 10$ consecutive subsets of observations that the model acquires recursively. Red elements represent the GP predictive posterior over the newer input domain while the blue ones refer to the past visited input space. Train and test data samples are plotted as colored crosses and dots respectively. Black crosses indicate the position of the inducing inputs at each time-step. The *magenta* line corresponds to the limit between the past and the new input domain explored by the continual GP.

value of 13.29×10^{-2}, which is slightly less than the initial fitting precision obtained at the time that the collection was first observed. Initially, the model obtained 13.13×10^{-2}. This means that, although the continal variational approach suffers a small decay in the predictive precision once past training samples are never revisited, the accuracy w.r.t. the test data remains constant *nine* steps after. This is, the GP has been rebuilt *nine* times via *nine* optimization processes where the uncertainty metrics are nor overwritten and neither forgotten.

II). Overlapping – In this version of the single-output experiment, we study potential difficulties of the GP regression model to accept overlapping partitions in a sequential manner. Here, we refer to overlapping partitions as the case where a few samples in the sequential collection revisit the previosly observed input space. As in the previous experiment, we consider sequences of $t = 10$ batches, and the sparse GP approximation is initialized with $M = 4$ points instead. We also limited the learning optimizer to a maximum of $t = 100$ iterations per run and importantly, the initial step of the continual process is trained using the standard variational bound of Hensman et al. (2015a); Saul et al. (2016); Moreno-Muñoz

Table 3.3: i) Streaming single-output data. Test-NLPD metrics ($\times 10^{-2}$). (COLUMN NEW BATCH) Predictive error values obtained in the new observed input area at each time-step ($t' = t$). (COLUMNS OLD BATCH) Predictive error values obtained in the past observed input areas at time-steps ($t' = 1, t' = 4$ and $t' = 8$). Colored values correspond to the GP prediction on the same test-samples at the t-th iteration. (COLUMN GLOBAL) NLPD values over the test-samples all along the input domain at each time-step t.

step	NEW BATCH $t' = t$	OLD BATCH $t' = 1$	OLD BATCH $t' = 4$	OLD BATCH $t' = 8$	**global**
$t = 1$	13.13 ± 0.10	-	-	-	13.13 ± 0.13
$t = 2$	12.50 ± 0.13	13.24 ± 0.10	-	-	25.74 ± 0.23
$t = 3$	12.54 ± 0.08	13.29 ± 0.13	-	-	38.48 ± 0.27
$t = 4$	11.59 ± 0.04	13.33 ± 0.12	-	-	52.26 ± 0.28
$t = 5$	11.34 ± 0.05	13.28 ± 0.10	11.34 ± 0.06	-	63.78 ± 0.32
$t = 6$	11.56 ± 0.06	13.29 ± 0.11	11.33 ± 0.06	-	75.35 ± 0.46
$t = 7$	12.71 ± 0.09	13.29 ± 0.12	11.34 ± 0.08	-	88.09 ± 0.55
$t = 8$	11.92 ± 0.05	13.29 ± 0.13	11.34 ± 0.06	-	100.01 ± 0.62
$t = 9$	13.55 ± 0.08	13.29 ± 0.09	11.34 ± 0.08	11.98 ± 0.06	113.60 ± 0.58
$t = 10$	11.73 ± 0.06	13.30 ± 0.14	11.34 ± 0.07	11.97 ± 0.04	125.34 ± 0.68

et al. (2018).

In Tab. 3.4, we show similar NLPD results to the ones included above in Tab. 3.3. The first column corresponds to the NLPD metrics obtained over the recently new samples at each tth time-step. Intermediate columns show the predictive GP accuracy over the past visited data. The last column represents the *global* fitting value, which augments as more test data is added to the evaluation.

Robustness to Propagation Errors

Additionally, we are particularly interested in the demonstration of the effect that the continual GP update has on the whole model. This is, how robust the sparse approximation can be as $t \to \infty$. Typically, the introduction of variational posterior densities $q(\cdot)$ as new prior beliefs into a Bayesian online scheme seems the most natural strategy to treat sequential observations using approximation methods. However, this approach is usually discarded in the literature due to the assumption that repetitive approximations may accumulate errors as the number of time-steps increases (Nguyen et al., 2018), something that usually happens with other probabilistic models. One of the main objectives in this experimental section is to beat this assumption, performing continual variational learning for signal processing applications with thousands of instances.

SOLAR PHYSICS DATA – Based on filtering experiments for signal processing applications, we obtained an astrophysics dataset which consists of the monthly average of *sunspot counting* numbers from 1700 to 1995. Particularly, we used the observations made for the analysis of sunspot cycles by the Royal Greenwich Observatory (US)[2]. For avoiding the use of non-tractable likelihood densities in the GP model, we transformed the strictly positive numbers into the real domain via the mapping $\log(1 + \boldsymbol{y})$. Our primary goal is to demonstrate that the predictive process of the continual model remains stable as $t \to \infty$, all over the input

[2]Solar physics data is publicly available at https://solarscience.msfc.nasa.gov/.

Table 3.4: ii) Overlapping single-output data. Test-NLPD ($\times 10^{-2}$). (COLUMN NEW BATCH) Predictive error values obtained in the new observed input area at each time-step ($t' = t$). (COLUMNS OLD BATCH) Predictive error values obtained in the past observed input areas at time-steps ($t' = 1, t' = 4$ and $t' = 8$). Colored values correspond to the GP prediction on the same test-samples at the t-th iteration. (COLUMN GLOBAL) NLPD values over the test-samples all along the input domain at each time-step t. In this experiment, input areas are overlapped with the previous one.

step	NEW BATCH $t' = t$	OLD BATCH $t' = 1$	OLD BATCH $t' = 4$	OLD BATCH $t' = 8$	**global**
$t = 1$	13.26 ± 0.29	-	-	-	13.26 ± 0.29
$t = 2$	11.70 ± 0.20	12.23 ± 0.10	-	-	23.94 ± 0.30
$t = 3$	13.60 ± 0.12	12.26 ± 0.11	-	-	37.58 ± 0.31
$t = 4$	12.63 ± 0.13	12.08 ± 0.17	-	-	50.37 ± 0.50
$t = 5$	14.50 ± 0.36	12.07 ± 0.12	12.66 ± 0.11	-	64.93 ± 0.77
$t = 6$	13.68 ± 0.16	12.04 ± 0.07	12.77 ± 0.10	-	79.38 ± 0.63
$t = 7$	13.80 ± 0.10	12.24 ± 0.09	12.75 ± 0.12	-	92.86 ± 0.73
$t = 8$	13.45 ± 0.09	12.03 ± 0.09	12.67 ± 0.11	-	106.21 ± 0.93
$t = 9$	12.64 ± 0.09	12.09 ± 0.08	12.69 ± 0.06	13.78 ± 0.09	119.04 ± 1.01
$t = 10$	12.84 ± 0.15	12.08 ± 0.11	12.71 ± 0.08	13.65 ± 0.09	131.93 ± 1.01

space, e.g. it does not forget past visited regions. In Fig. 3.12, we show three captures of the continual learning process until a maximum of $t = 10^3$ iterations. We remark that we used a one-sample update rule, so the experiment consisted of 10^3 consecutive optimization trials. For preserving tractability, we setup an initial number of $M = 10$ inducing inputs for the *warm up* period and later it was increased with one new point every 100 new samples. A demonstrative visualization of the whole experiment can be found at https://www.youtube.com/watch?v=j7kpru4YrcQ.

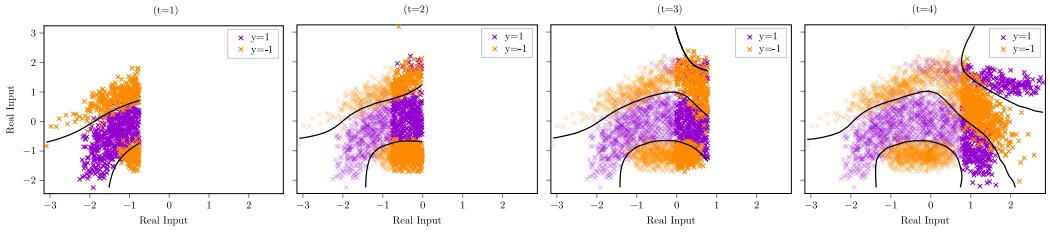


Figure 3.11: Performance of the continual GP learning approach under non-Gaussian data for binary classification tasks. Past samples are plotted in a grey scaled version. Black curves represent the frontier between positive and negative predictions w.r.t. the output values. Additionally, the last r.h.s. plot shows the final prediction of the model over the entire 2-dimensional input space, within the last training data seen so far (sharp colors).

Continual GP Classification

Since the approach presented in this paper is also valid under the presence of non-Gaussian likelihood distributions, we introduced an additional experiment for binary data. Particularly, we chose the *banana* dataset, widely used for demonstrative results of scalable GP

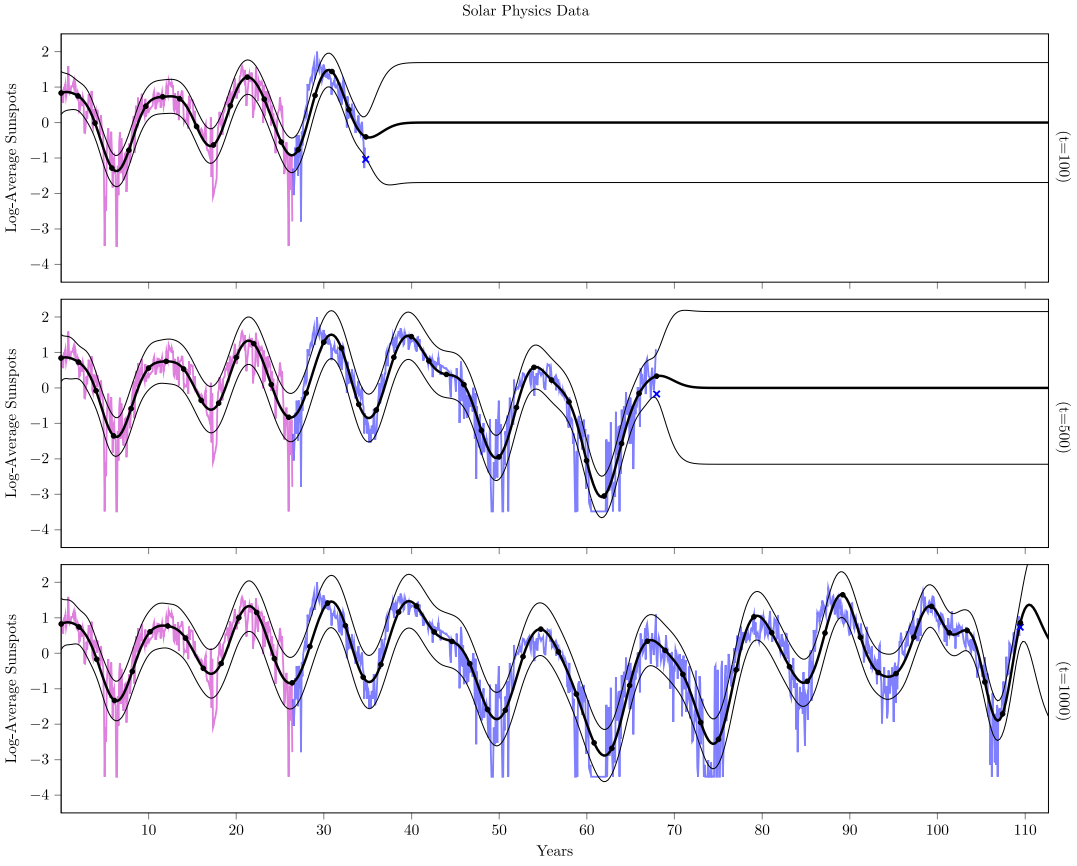


Figure 3.12: Results for single-output regression on solar physics data with one-sample updates of the continual sparse GP model. Pink colored signal corresponds to the *warm up* observations in the batch mode. Greyed blue signals correspond to the former visited observations while the blue cross is the new incoming one. Black colored curves correspond to the mean function and the 95% confidence interval of the predictive GP distribution all over the input-space, computed at each time iteration. Black dots are the inducing variables at each time-step.

classification tasks (Hensman et al., 2015a; Bui et al., 2017a).

The banana dataset consists of $N = 5200$ pairs of input-output observations, where we selected a 30% for testing the predictive error metrics and the rest for training. All input objects have a dimensionality $p = 2$. The experiment was divided into four steps of inference, depicted in Fig. 3.11. The final NLPD and error rates (ER) are provided in the appendix. We can see that the performance of the continual method is similar to standard GP classification. The precision remains constant in the input areas where training data is never revisited.

Continual MOGP Channels

As we discussed in Sec. 3.2.2, the MOGP framework introduces two *layers* of inference. One is linked to the LFs $\mathcal{U}$, where the sparse approximation lies, while the other focus in the observations and the parameterization of their likelihood densities via $\mathcal{F}$. This two-layer structure makes the continual multi-output learning process work in a different way w.r.t.

the marginal lower bound. Now, the expectation terms are *decoupled* from the regularization, which is only focused on the distributions over the LFs. The key property of this approach is that we may consider extremely irregular problems, where for instance output sequences are completely assymetric. An example is shown in Fig. 3.5. Two experiments with both real-world and synthetic data are provided in the appendix.

3.4 Discussion

We have described novel inference methods for GP models under the presence of sequential data and distributed scenarios. In the continual case, we have presented an approach that extends the existing posterior-prior recursion of online Bayesian inference to the infinite-dimensional framework of GP models. The key principle behind is the reconstruction of implicit GP priors over the space-of-functions conditioned to the past posterior densities via the GP predictive equation. The model is adapted for sparse approximation and variational methods as well as we applied it to the MOGP setting. Additionally, we introduced a second GP framework for building global approximations from distributed GP models. The main contribution of this approach is the construction of the ensemble bound which accepts tasks from regression, classification and heterogeneous problems. Experimental results show evidence that both statistical methods are robust in the scenarios considered. Online inference is also revisited in the next Ch. 4, where we discuss its implications for change-point detection problems.

CHAPTER 4

CHANGE-POINT DETECTION

ONCE a probabilistic model is learned in a data-driven manner, the problem of detecting *changes* or *anomalies* given that model and its adjacent parameters appears in many scientific disciplines, e.g. ECG/EEG analysis (Agudelo-España et al., 2020), DNA segmentation (Braun et al., 2000), spatio-temporal modeling (Knoblauch and Damoulas, 2018) or econometrics (Chen and Gupta, 1997). The traditional question use to be if newer observations conform or not to the expected behavior indicated by the already fitted model. Moreover, the problem can be looked from two perspectives that are interrelated, one more oriented to the characterisation of a single observation as an *outlier*, typically named as *anomaly detection*, and another focused in refreshing a learning model that no longer fits to the sequence of new data. This last problem, widely known as *change-point detection* (CPD) is the main core of this chapter of the thesis.

• A formal definition of an *anomaly* would be a single observation that does not fit with the rest of data and hence, to the model.

We begin in Sec. 4.1 by reviewing the principles behind CPD methods, and in particular, the probabilistic approaches based on the Bayesian perspective. After analysing the main drawbacks that this sort of solutions may lead to. In Sec. 4.2, we present a novel extension within latent variable models, named *hierarchical* CPD. The experimental results on this side remark the utility of locating the change-point problem in a low-dimensional manifold where performing Bayesian inference is significantly easier.

Once the hierarchical version of Bayesian CPD is presented, two more extensions of the approach are developed. In Sec. 4.2.4, we consider the case of unfixing the dimensionality of the discrete latent variables, that is, let it increase until a potentially infinite number of classes. This solution is strongly related to *Bayesian nonparametrics* (BNP) (Antoniak, 1974), and more in detail, to the well-known *chinese restaurant process* (CRP) (Pitman, 2002). Whilst the infinite hierarchical CPD model is included in this chapter, the general idea for inference in a continual learning setup is also motivated by the framework described in the preceding Ch. 3. As we describe in Sec. 4.2.5, the sequential inference of the surrogate latent variable model can also be improved. Hence, we increase the robustness of the hierarchical model by introducing a novel methodology based on a simple multinomial sampling method.

The results described in this chapter were presented in three main pieces of work. The first one, Moreno-Muñoz et al. (2018), was accepted in the Pattern Recognition Journal and it is currently awaiting for its formal publication. The second one, Moreno-Muñoz et al. (2020b), was presented at the IEEE 2020 International Conference on Acoustics, Speech and Signal Processing (ICASSP). Finally, in collaboration with L. Romero-Medrano, Romero-Medrano et al. (2020), was presented in the 2020 IEEE International Workshop on Machine Learning for Signal Processing (MLSP) as an oral presentation.

4.1 Bayesian Change-point Detection

Change-point detection (CPD) usually refers to the problem of locating abrupt transitions in the generative model of a sequence of data observations. Detecting these abrupt transitions or change-points (CPs) has been long studied and appears in a vast amount of real-world

scenarios. To name a few, for instance, investment strategies, the analysis of social networks or cognitive radio in signal processing.

The problem of CPD is typically faced from two statistical perspectives, one probabilistic and another more oriented to frequentist methods. Both of them usually require a model defined *a priori*, or alternatively, the introduction of model-free methods. The main idea behind frequentist methods is to derive a measure of discrepancy, often based on likelihood ratio tests (Kuncheva, 2011), between the pre-change and post-change distributions. Once the values of discrepancy are estimated, these are compared to some threshold value and a decision is taken accordingly. On the other hand, probabilistic approaches, and particularly those ones based on the Bayesian principles aim to assign a prior distribution over the CPs, or a proper surrogate, and derive a posterior distribution given the data. This last strategy is the one that we analyze and expand in this chapter.

Short Review of Bayesian CPD Methods

Typical approaches to CPD focus on batch settings (Harchaoui et al., 2009; Fearnhead, 2006), where the number of partitions is often unknown. In contrast, two *online* CPD methods were introduced, via particle filtering (Fearnhead and Liu, 2007) and conjugate exponential-families (Adams and MacKay, 2007). In the *online* case, the data are obtained incrementally over time, and inference updates are required each time that a new object is observed. Both works infer the location of change-points for univariate time-series using MAP estimation. A different approach is developed in Li et al. (2015), where the changes are detected in sequences whose time-horizon can be either fixed or not. Additionally, there has been a considerable effort to apply these methods on different statistical domains. For instance, Höhle (2010) introduced an online cumulative-sum detector that finds changes in categorical sequential data. Other approaches are able to model multivariate collections using undirected graphical models (Xuan and Murphy, 2007).

Following the Bayesian perspective, several extensions of the Bayesian online CPD (BOCPD) model of Adams and MacKay (2007) have been developed. These include adaptive sequential methods (Turner et al., 2009) and Gaussian processes (Saatçi et al., 2010). This last work extends the BOCPD algorithm to locate change-points from observations with an arbitrary temporal structure. Moreover, unlike previous methods, the non-exponential CPD model in Turner et al. (2013) explored the application of BOCPD to new families of distributions, where computing posterior probabilities is intractable and variational inferences is therefore required. However, there is a general lack of methods that address the problem of CPD in heterogeneous data collections where observations might be of different statistical data types.

Besides the heterogeneity problem, there is another key challenge in the state-of-the-art to scale up CPD for high-dimensional data. This has not been considered much in the past. One exception is the work of Xie et al. (2013), which considers that data belongs to a low-dimensional manifold embedded in some high-dimensional ambient space. A similar work based on introducing latent structure into the Bayesian model is Agudelo-España et al. (2020), which uses HMMs. These last ideas are similar to the ones presented in this section since they reduce the computational cost by focusing the change-point analysis to some latent space of interest.

4.1.1 Bayesian Online Change-Point Detection

We start by considering a time series, with observations $\boldsymbol{x}_{1:t} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t\}$, that are divided into non-overlapping partitions. Each partition is denoted by ρ_i with $i = 1, 2, \ldots$ Since the division of such partitions is assumed to be non-overlapping, we assume that a

• At this point, we avoid to define the natural domain of $\boldsymbol{x}_t$ as we will later generalize the solution for both continuous and discrete random variables.

CP is the point between two of them. Based on Adams and MacKay (2007), we assume that the data within each partition ρ_i is independent and identically distributed (i.i.d.) according to some generative probability distribution $p(\boldsymbol{x}_t|\boldsymbol{\theta}_{\rho_i})$, where the parameter vector $\boldsymbol{\theta}_{\rho_i}$ is unknown *a priori*. Under this assumption, one may determine the change-points in an intuitive manner as the changes in the parameters:

$$\boldsymbol{\theta}_t = \begin{cases} \boldsymbol{\theta}_{\rho_1}, & t < \text{CP}_1, \\ \boldsymbol{\theta}_{\rho_2}, & \text{CP}_1 \leq t \leq \text{CP}_2, \\ \boldsymbol{\theta}_{\rho_3}, & \text{CP}_2 \leq t \leq \text{CP}_3, \\ & \vdots \end{cases} \tag{4.1}$$

The main idea introduced in Adams and MacKay (2007), is the *run-length* r_t, which is defined as a discrete random variable that counts the number of time-steps since the last CP. That is,

$$r_t = \begin{cases} 0, & \text{CP at time } t, \\ r_t + 1, & \text{otherwise}, \end{cases} \tag{4.2}$$

and might be seen as a proxy for the true position for change-points. The general objective of the Bayesian online change-point detection (BOCPD) algorithm is to recursively compute the posterior distribution over such *run-length* $p(r_t|\boldsymbol{x}_{1:t})$, from which ones aims to identify a CP if the probability mass accumulates near $r_t = 0$. That is, the number of time-steps since the last CP is almost zero and hence, a new partition is assumed to begin.

The posterior distribution $p(r_t|\boldsymbol{x}_{1:t})$ is directly obtained by marginalization of the discrete values of r_t seen so far at the time-step t in the joint distribution $p(r_t, \boldsymbol{x}_{1:t})$. The computation of this last joint distribution is indeed obtained from the following recursive factorization,

$$\begin{aligned} p(r_t, \boldsymbol{x}_{1:t}) &= \sum_{r_t - 1} p(r_t, r_{t-1}, \boldsymbol{x}_{1:t}) \\ &= \sum_{r_t - 1} p(r_t|r_{t-1}) p(\boldsymbol{x}_t, r_{t-1}, \boldsymbol{x}_{1:t-1}) \\ &= \sum_{r_t - 1} p(r_t|r_{t-1}) p(\boldsymbol{x}_t|r_{t-1}, \boldsymbol{x}_{1:t-1}) p(r_{t-1}, \boldsymbol{x}_{1:t-1}), \end{aligned} \tag{4.3}$$

where the conditional *run-length* prior $p(r_t|r_{t-1})$ is given by

$$p(r_t|r_{t-1}) = \begin{cases} h(r_{t-1} + 1), & r_t = 0, \\ 1 - h(r_{t-1} + 1), & r_t = r_{t-1} + 1, \end{cases} \tag{4.4}$$

and $h(\cdot)$ is the *hazard* function that we consider to be constant along time with a given time-scale hyperparameter τ, also fixed. On the other hand, the posterior predictive distribution $p(\boldsymbol{x}_t|r_{t-1}, \boldsymbol{x}_{1:t-1})$ is perhaps the most important term for us in the recursive factorization. Given the *underlying* predictive model (UPM) in the change-point detection method, that is, having established a likelihood function of the form $p(\boldsymbol{x}_t|\boldsymbol{\theta}_t)$, one exploits the update of parameters $\boldsymbol{\theta}_t$ to see if the new datum $\boldsymbol{x}_t$ fits or not to the previous inferred model. This process is naturally handled by the predictive integral

$$p(\boldsymbol{x}_t|r_{t-1}, \boldsymbol{x}_{1:t-1}) = \int p(\boldsymbol{x}_t|\boldsymbol{\theta}_t) p(\boldsymbol{\theta}_t|r_{t-1}, \boldsymbol{x}_{1:t-1}) d\boldsymbol{\theta}_t, \tag{4.5}$$

whose predictive values are typically set at each time-step t as $\Psi_t^{(r)}$. To clarify this point, the parameter estimation and predictive mechanism are depicted in Fig. 4.1, where several

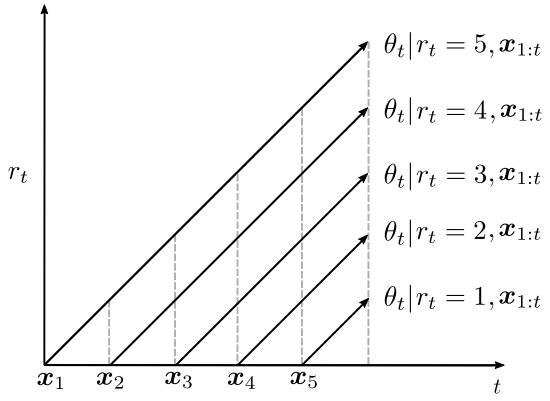


Figure 4.1: Illustration of the parallel inference mechanism for the recursive estimation of $\boldsymbol{\theta}_t$ conditioned on the run-length r_t given the sequence $\boldsymbol{x}_{1:t}$.

threads of inference work at the same time. The conditioning on the *run-length* variable r_t implies that a new observation $\boldsymbol{x}_t$ will be evaluated for each parallel thread, whose parameters $\boldsymbol{\theta}_t|r_t, \boldsymbol{x}_{1:t}$ are slightly different.

• We often refer to the *threads* of inference as the process of estimating parameters $\boldsymbol{\theta}_t$ from different subsets of data in the sequence. A potential synonym would be just to say *partition hypotheses*.

In the original idea of Adams and MacKay (2007), the estimation of $p(\boldsymbol{\theta}_t|r_{t-1}, \boldsymbol{x}_{1:t-1})$ distributions was carried out via exponential family likelihoods (Wainwright and Jordan, 2008) and their corresponding *conjugate priors*. Under this convenient form of recursive inference, only a small number of *sufficient statistics* from each new data point $\boldsymbol{x}_t$ are necessary to update the posterior estimates. However, in those cases where the use of conjugate systems is not possible, the problem of recursive inference worsens.

Once we obtain the joint distribution $p(r_t, \boldsymbol{x}_{1:t})$, the posterior over r_t can be directly found as

$$p(r_t|\boldsymbol{x}_{1:t}) = \frac{p(r_t, \boldsymbol{x}_{1:t})}{p(\boldsymbol{x}_{1:t})}, \tag{4.6}$$

where we must compute the marginal likelihood or evidence as $p(\boldsymbol{x}_{1:t}) = \sum_{r_t} p(r_t, \boldsymbol{x}_{1:t})$. Additionally, if the *run-length* variable is not of interest for the full inference process in a given scenario, it also can be marginalized. This point lead us to obtain a predictive distribution that is no longer conditioned to the CP hypothesis. To find the marginal predictive distribution, we just marginalize over r_t as

$$p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_{1:t}) = \sum_{r_t} p(\boldsymbol{x}_{t+1}|r_t, \boldsymbol{x}_{1:t}) p(r_t|\boldsymbol{x}_{1:t}), \tag{4.7}$$

which might be of interest for *online* prediction when the underlying task is always changing or the task segmentation is not the final goal, e.g. in the case of continuous meta-learning (Harrison et al., 2019) problems.

High-Dimensional Issues

A natural problem arises in Bayesian CPD when the observations $\boldsymbol{x}_t$ become more and more complex, i.e. high-dimensional and/or of different statistical data-types. The first intution tell us that larger a datum $\boldsymbol{x}_t$ is, higher is the number of $\boldsymbol{\theta}_t$ parameters needed. Due to the recursive nature of the posterior $p(r_t|\boldsymbol{x}_{1:t})$ computation, the BOCPD algorithm may be unable to accumulate enough probability mass on low values of r_t when the time between CPs becomes of the order of magnitude (or less) of the number of model parameters $\boldsymbol{\theta}_t$.

This fact makes almost impossible to detect CPs in a reliable manner from high-dimensional observations.

From a simpler inference perspective, the problem of high-dimensionality in Bayesian CPD is similar to other scenarios in probabilistic machine learning. As the number of parameters $\boldsymbol{\theta}_t$ needed for modelling the data $\boldsymbol{x}_{1:t}$ increases, the *curse of dimensionality* (Bellman, 1961) forces us to observe more samples to be certain on our posterior discoveries. However, in a sequential setup as the one considered in Adams and MacKay (2007), the observation of a higher order of data samples is no longer possible as the position of each CP is fixed at some time-step t. As a consecuence, the posterior distribution $p(\boldsymbol{\theta}_t|\boldsymbol{x}_{1:t})$ will be of higher uncertainty, leading to noisy $\Psi_t^{(r)}$ or predictive probabilities. In the end, the whole CPD process will tend to fail in the detection as the number of parameters increases and the number of data points (hence information) between CPs cannot be augmented.

4.2 Hierarchical Change-Point Detection

The problem of high-dimensional data in Bayesian CPD, and particularly within the BOCPD algorithm, leads to complex generative models can be overcome by introducing hierarchy into the model. To overcome the limitation imposed by the large number of parameters $\boldsymbol{\theta}_t$, one may assume that even if the data are high-dimensional, they belong to a low-dimensional manifold as follows

$$p(\boldsymbol{x}_t|\boldsymbol{\theta}_t) = \int p(\boldsymbol{x}_t|\boldsymbol{z}_t)p(\boldsymbol{z}_t|\boldsymbol{\theta}_t)d\boldsymbol{z}_t, \tag{4.8}$$

where $\boldsymbol{z}_t$ is the vector of latent variables corresponding to the observation $\boldsymbol{x}_t$. An important assumption is that parameters that change between change-points now parameterize the generative distribution $p(\boldsymbol{z}_t|\boldsymbol{\theta}_t)$ of the latent variable. Since the dimensionality of $\boldsymbol{z}_t$ is designed to be significantly smaller than the one of $\boldsymbol{x}_t$, then the total size of $\boldsymbol{\theta}_t$ is also reduced.

A preliminary idea is to consider that the latent variable $\boldsymbol{z}_t$ may be either continuous or discrete, and the conditional distribution $p(\boldsymbol{x}_t|\boldsymbol{z}_t)$ is assumed to be fixed and known *a priori*. Given this sort of hierarchical model, we are still interested in the *run-length* variable r_t and its posterior distribution $p(r_t|\boldsymbol{x}_{1:t})$, which in turn will indicate us the location of each CP. To appropiately adapt the problem to the underlying latent sequence $\boldsymbol{z}_{1:t}$ in the recursive factorization of Eq. (4.3), we should look for the joint distribution $p(r_t, \boldsymbol{x}_{1:t}, \boldsymbol{z}_{1:t}, \boldsymbol{\theta}_t)$. It can be also reduced to the result in Eq. (4.6) as

$$p(r_t, \boldsymbol{x}_{1:t}) = \iint p(r_t, \boldsymbol{x}_{1:t}, \boldsymbol{z}_{1:t}, \boldsymbol{\theta}_t)d\boldsymbol{\theta}_t d\boldsymbol{z}_{1:t}. \tag{4.9}$$

For any familiar reader with latent variable models in sequential problems, this procedure can seem quite involved. Mainly, due to the integrals in Eq. (4.9) are typically difficult to compute and it is not trivial to achieve a recursive expression as in Adams and MacKay (2007). However, to obtain the required recursivity, we know that the joint distribution $p(r_t, \boldsymbol{x}_{1:t}, \boldsymbol{z}_{1:t}, \boldsymbol{\theta}_t)$ must factorise as

$$p(r_t, \boldsymbol{x}_{1:t}, \boldsymbol{z}_{1:t}, \boldsymbol{\theta}_t) = p(\boldsymbol{x}_{1:t}|\boldsymbol{z}_{1:t})p(r_t, \boldsymbol{z}_{1:t}, \boldsymbol{\theta}_t), \tag{4.10}$$

where we have exploited the previous assumption where the conditional $p(\boldsymbol{x}_t|\boldsymbol{z}_t)$ is fixed and known. Perhaps the easiest point of this derivation is the marginalisation of parameters $\boldsymbol{\theta}_t$ in Eq. (4.9), which yields

$$p(r_t, \boldsymbol{z}_{1:t}, \boldsymbol{x}_{1:t}) = p(\boldsymbol{x}_{1:t}|\boldsymbol{z}_{1:t})p(r_t, \boldsymbol{z}_{1:t}). \tag{4.11}$$

A priori, it seems obvious that the recursive nature of the algorithm is still there, as a direct consequence of the factorization in Eq. (4.3), that leads to

$$p(r_t, \boldsymbol{z}_{1:t}, \boldsymbol{\theta}_t) = \sum_{r_t - 1} p(r_t|r_{t-1})p(\boldsymbol{z}_t|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t|r_{t-1}, \boldsymbol{z}_{1:t-1})p(r_{t-1}, \boldsymbol{z}_{1:t-1}), \tag{4.12}$$

where the sequence $\boldsymbol{z}_{1:t-1}$ is now part of the recursive predictive mechanism depicted in Fig. 4.1 instead of the previous observations. Importantly, the last term in the previous expression is also a result of integrating the parameters, and can be obtained as

$$p(r_{t-1}, \boldsymbol{z}_{1:t-1}) = \int p(r_{t-1}, \boldsymbol{z}_{1:t-1}, \boldsymbol{\theta}_{t-1}) d\boldsymbol{\theta}_{t-1}, \tag{4.13}$$

and indeed, would lead to the same expression of Eq. (4.3), where the posterior predictive distribution $p(\boldsymbol{z}_{t-1}|r_{t-1}, \boldsymbol{z}_{1:t-2})$ lies on. Having said this, and under the combination of all previous results, the probability $p(r_t, \boldsymbol{z}_{1:t})$ of interest also factorizes and becomes

$$p(r_t, \boldsymbol{z}_{1:t}) = \sum_{r_t - 1} p(r_t|r_{t-1}) \Psi_t^{(r)} p(r_{t-1}, \boldsymbol{z}_{1:t-1}), \tag{4.14}$$

where the predictive over the underlying predictive model focuses now on the latent counterpart, that is

$$\Psi_t^{(r)} = p(\boldsymbol{z}_t|r_{t-1}, \boldsymbol{z}_{1:t-1}) = \int p(\boldsymbol{z}_t|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t|r_{t-1}, \boldsymbol{z}_{1:t-1}) d\boldsymbol{\theta}_t. \tag{4.15}$$

So far, we have preserved the recursive mechanism of the original BOCPD but adapted to the introduction of latent variable models. However, we might face the first problem in Eq. (4.9) due to the marginalizaion of $\boldsymbol{\theta}_t$ at each time-step. Notice that the exact evaluation of $\Psi_t^{(r)}$ is often intractable when the underlying latent variable model chosen is non-Gaussian. Eventually, this may require to introduce approximate or numerical methods, which in turn reduces the efficiency of the CPD algorithm.

The second problem that we must overcome is the marginalization over the unknown sequence of latent variables $\boldsymbol{z}_{1:t}$ in Eq. (4.9). To address this point, we propose to use a simple latent *class* model, where the variables are given by $z_t \in \{1, 2, \ldots, K\}$, with K being the total number of classes. Moreover, this choice also results in a small number of unknown parameters $\boldsymbol{\theta}_t$ and hence, a better performance of the hierarchical CPD algorithm.

The introduction of latent class model also has an additional purpose, that we previously developed in Chap. 2. The use of an hierarchical approach in CPD simplifies the interpretability of the algorithm under the presence of high-dimensional, heterogeneous or missing data. To sum up, the observation $\boldsymbol{x}_t$ has now a single discrete representation z_t, which indicates the *true* latent class that it belongs to. The general objective now to perform CPD is to infer the *univariate* sequence $\boldsymbol{z}_{1:t} = [z_1, z_2, \ldots, z_t]^\top$ of indicators.

Sequential Marginalization of Latent Variables

An intuitive strategy in probability-based models is to marginalize those variables that are uncertain or simply difficult to infer given the data. In the particular case of hierarchical CPD, where the secuence of latent variables $\boldsymbol{z}_{1:t}$ is introduced to reduce the parametric complexity of the model, we aim to marginalize out every z_t in a sequential manner. However, due to the recursive dependence in Eq. (4.14), marginalizing the whole sequence of latent class variables yields a hard combinatorial problem, which can be computationally challenging.

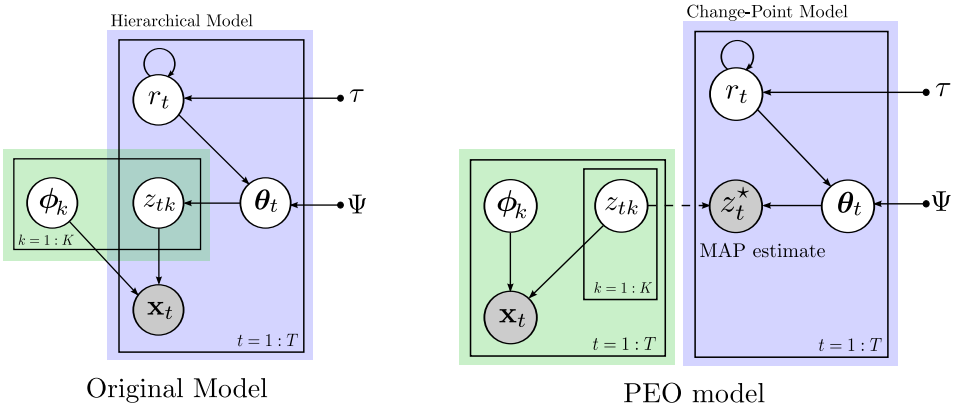


Figure 4.2: Graphical representation of the original (LEFT) and simplified (RIGHT) hierarchical models. Shaded nodes correspond to observed variables and the dashed line denotes how the posterior probabilities of z_t are observed by the change-point model (blue). The variable, $z_t^\star$ is a point estimate which takes the z_{tk} value with highest probability. Additionally, Ψ denotes the fixed hyperparameters from the corresponding priors placed on $\boldsymbol{\theta}_t$.

Concretely, for a sequence $\boldsymbol{z}_{1:t}$ of length T, with K classes, the explicit evaluation of $p(r_t, \boldsymbol{x}_{1:t})$ requires $\mathcal{O}(K^T)$ operations. Thus, it may not be possible to compute it for long observation periods T, and/or large number of classes K.

In the next sections of this chapter, we develop a solution for the hierarchical CPD model that overcomes the problem of a computational complexity that grows exponentially fast as T and K increases.

4.2.1 Point-Estimate Observations

Having introduced an hierarchical model in the original BOCPD algorithm of Adams and MacKay (2007), we now face the problem of computational complexity. Among the possible latent variable models to consider, we propose to use latent classes z_t, such that the generative distribution of our i.i.d. data can be rewritten as

$$p(\boldsymbol{x}_t|\boldsymbol{\theta}_t) = \sum_{z_t=1}^{K} p(\boldsymbol{x}_t|z_t)p(z_t|\boldsymbol{\theta}_t), \tag{4.16}$$

where z_t is a categorical r.v. and K its dimensionality. This sort of latent class model can be also seen as mixture model. As we have shown, under the presence of such latent variables, the natural recursive integration of the CPD method becomes unfeasible very quickly as T and K increase. However, we are still interested in maintaining such latent representation z_t in the context of change-points. In the previous Sec. 4.2, we saw that the main factor of this complexity is the sum all over the combinations of $\boldsymbol{z}_{1:t-1}$.

To avoid the marginalization while keeping the latent variable methodology, we simplify the previous hierarhical model by considering an alternative strategy. If collapsing the entire latent domain is too costly, we propose to *observe* it (Nazábal et al., 2016). That is, we can explicitly observe point-estimates of z_t, which were hidden before. In particular, instead of marginalizing the sequence $\boldsymbol{z}_{1:t}$, we directly *plug in* the point-estimate values taken by z_t at each time-step.

To adapt the hierarchical model to the new *pseudo-observed* latent variables, we take the set of point-estimates $\boldsymbol{z}^{\star}_{1:t}$ as the input data to the detector. We further assume the likelihood distribution to be $p(\boldsymbol{z}^{\star}_t|\boldsymbol{\theta}_t)$, that is, we are directly modelling the changes on the sequence of point-estimates. Particularly, these variables are given by the *maximum a-posteriori* (MAP) criterion. That is,

$$z^{\star}_t = \arg\max_{z_t} p(z_t|\boldsymbol{x}_t), \tag{4.17}$$

for which we need to previously derive the posterior distribution $p(z_t|\boldsymbol{x}_t)$. In this chapter, we assume that it is given as long as we focus in the particular task of CPD. To obtain this posterior distribution, we presented the latent variable models for high-dimensional, heterogeneous and corrupted data in Chap. 2 and several alternatives for inference in the previous Chap. 3. The graphical model included in Fig. 4.2 shows the complete details of this *point-estimate observation* (PEO) approximation.

To perform hierarchical CPD on the simplified version of the algorithm, we first choose the likelihood function for $p(z^{\star}_t|\boldsymbol{\theta}_t)$ to be a categorical distribution with natural parameter $\boldsymbol{\pi}_t$. We also place a Dirichlet prior on $\boldsymbol{\pi}_t$ with a single hyperparameter $\boldsymbol{\gamma}$ for preserving the conjugacy. Then, the generative model turns to be $z^{\star}_t \sim \text{Cat}(\boldsymbol{\pi}_t)$ and $\boldsymbol{\pi}_t \sim \text{Dirichlet}(\boldsymbol{\gamma})$, where $\boldsymbol{\pi}_t \in \mathcal{S}^K$ and $\boldsymbol{\gamma} \in \mathbb{R}^K_+$, with $\mathcal{S}^K$ and $\mathbb{R}^K_+$ being the K-dimensional simplex and the positive orthant, respectively. Interestingly, this choice for the prior distributions allows us to still compute the predictive probabilities $\pi^{(r)}_t$ in a closed-form manner, which are given by

$$\pi^{(r)}_t = p(z^{\star}_t|r_{t-1}, \boldsymbol{z}^{\star}_{1:t-1}) = \frac{\gamma^{(r)}_k}{\sum_{k'=1}^{K}\gamma^{(r)}_{k'}}, \ \forall r \in \{1, 2, \ldots, t\}, \tag{4.18}$$

where $\gamma^{(r)}_t$ is the kth component of the vector parameter $\boldsymbol{\gamma}_t$ of the Dirichlet prior distribution computed for the partition hypothesis indicated by the *run-length* r_t. The expression, which is a direct consequence of the Dirichlet-categorical conjugate system, provides a significant reduction in the computational complexity and results in a very simple method. Finally, the parameters are updated following the rule $\gamma^{(r)}_k \leftarrow \gamma^{(r)}_k + \mathbb{I}\{z^{\star}_t = k\}$, with $\mathbb{I}\{\cdot\}$ being the indicator function as in the latent-class models previously considered. The PEO model is effective in terms of computational cost, while at the same time, it provides a good performance. We depict an example with synthetically generated data in Fig. 4.3.

4.2.2 Sampling Alternatives for Approximate Inference

As a second approximation to the hierarchical model and also based on Nazábal et al. (2016), we may use the full vector of the posterior probabilities from latent variables, $p(z_t|\boldsymbol{x}_t)$. Similarly as Nazábal et al. (2016) does, multivariate *pseudo-observations* can be considered. That is, the approximation would take the new observable inputs $\tilde{\boldsymbol{z}}_t = p(z_t|\boldsymbol{x}_t)$, where $\tilde{\boldsymbol{z}}_t \in \mathcal{S}^K$ to satisfy $\sum_{k=1}^{K}\tilde{z}^k_t = 1$. The main advantage of this approach is that even if the *true* latent class z_t is unknown, we are able to detect CPs from the sequence of posterior probabilities. Intuitively, and similarly to the previous simplified model, the cost of marginalization is avoided. We denote this approximation as the *full posterior observation* (FPO) model.

The likelihood function should be $p(\tilde{\boldsymbol{z}}_t|\boldsymbol{\theta})$, which is taken to be a Dirichlet distribution with natural parameters $\eta\boldsymbol{\lambda}$. This *decoupled* parametric form simplifies the inference process to learn the inverse variance $\eta \in \mathbb{R}_+$ and the mean vector $\boldsymbol{\lambda} \in \mathcal{S}^K$. This decomposition allows us to choose a Gamma prior for η and a second Dirichlet distribution for $\boldsymbol{\lambda}$. Then, we may rewrite the probability model of the FPO approximation as

$$\tilde{\boldsymbol{z}}_t \sim \text{Dirichlet}(\eta\boldsymbol{\lambda}), \tag{4.19}$$

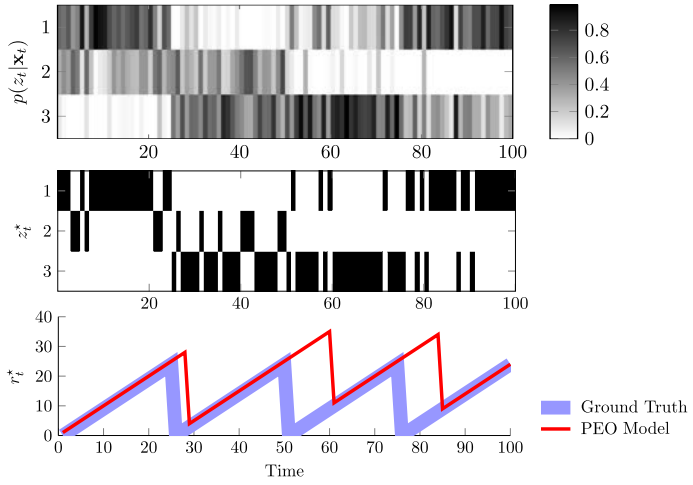


Figure 4.3: Performance of the PEO simplified model. (TOP ROW:) Sequence of posterior probability vectors $p(z_t|\mathbf{x}_t)$ where $K = 3$ and $T = 100$. Three change points are placed at $t = \{25, 50, 75\}$. (MIDDLE ROW:) Sequence of point estimates $z_t^\star = \arg\max_z p(z_t|\mathbf{x}_t)$ as 1-of-K encoding. (BOTTOM ROW:) Ground truth of change points (blue), CP estimation $r_t^\star$ from PEO model (red).

where $\eta \sim \text{Ga}(\kappa, \nu)$ and $\boldsymbol{\lambda} \sim \text{Dirichlet}(\boldsymbol{\beta})$, with $\boldsymbol{\beta} \in \mathcal{S}^K$ and $k, v \in \mathbb{R}_+$. Now, to obtain the joint distribution $p(r_t, \tilde{\boldsymbol{z}}_{1:t})$, similarly as we did in Eq. (4.14), we need to compute the predictive integral $\pi_t^{(r)}$. Notice that it is equivalent to the one presented in Eq. (4.15). To compute this integral, a strong requirement is to first obtain the posterior distribution $p(\boldsymbol{\theta}_t|r_{t-1}, \tilde{\boldsymbol{z}}_{1:t-1})$. Without the help of any conjugate exponential system, it takes the form

$$p(\boldsymbol{\theta}_t|r_{t-1}, \tilde{\boldsymbol{z}}_{1:t-1}) = p(\eta_t, \boldsymbol{\lambda}_t|r_{t-1}, \tilde{\boldsymbol{z}}_{1:t-1}) \propto p(\eta_t)p(\boldsymbol{\lambda}_t) \prod_{\tau=1}^{r_{t-1}} p(\tilde{\boldsymbol{z}}_\tau|\eta_\tau, \boldsymbol{\lambda}_\tau) \tag{4.20}$$

$$= \text{Ga}(\eta_t|\kappa, \nu)\text{Dir}(\boldsymbol{\lambda}_t|\boldsymbol{\beta}) \prod_{\tau=1}^{r_{t-1}} \text{Dir}(\tilde{\boldsymbol{z}}_\tau|\eta_\tau, \boldsymbol{\lambda}_\tau), \tag{4.21}$$

where we have used the two prior distributions defined above with fixed hyperparameters κ, ν and $\boldsymbol{\beta}$. Thus, the predictive integral that we look for, becomes

$$\pi_t^{(r)} = p(\tilde{\boldsymbol{z}}_t|r_{t-1}, \tilde{\boldsymbol{z}}_{1:t-1}) = \int p(\tilde{\boldsymbol{z}}_t|\eta_t, \boldsymbol{\lambda}_t)p(\eta_t, \boldsymbol{\lambda}_t|r_{t-1}, \tilde{\boldsymbol{z}}_{1:t-1})d\eta_t d\boldsymbol{\lambda}_t. \tag{4.22}$$

However, as we already pointed out, this integral is analytically *intractable*. Instead, we propose to solve it via Markov chain Monte-Carlo (MCMC) as follows

$$\pi_t^{(r)} \approx \frac{1}{S} \sum_{s=1}^{S} p(\tilde{\boldsymbol{z}}_t|\eta_t^s, \boldsymbol{\lambda}_t^s), \tag{4.23}$$

where $\{\eta_t^s, \boldsymbol{\lambda}_t^s\}_{s=1}^S$ are the corresponding samples of $\{\eta_t, \boldsymbol{\lambda}_t\}$, with S being the total number of samples. Particularly, we use a Gibbs sampler to draw realizations from the conditional distribution $p(\eta_t, \boldsymbol{\lambda}_t|r_{t-1}, \tilde{\boldsymbol{z}}_{1:t-1})$. The equations for the conditional probabilities are given

Algorithm 3 – Hierarchical CPD within FPO model

Input: Observe $\boldsymbol{x}_{1:t} \rightarrow$ obtain $\tilde{\boldsymbol{z}}_{1:t} = p(\boldsymbol{z}_{1:t}|\boldsymbol{x}_{1:t})$
for $\tilde{\boldsymbol{z}}_t$ **in** $\tilde{\boldsymbol{z}}_{1:t}$ **do**
  for $r_{t-1} = 1$ **to** t **do**
    Evaluate $\pi_t^{(r)} = \frac{1}{S}\sum_{s=1}^{S} p(\tilde{\boldsymbol{z}}_t|\boldsymbol{\lambda}_t^s, \eta_t^s)$
  end for
  Compute growth probabilities: $p(r_t = r_{t-1} + 1, \tilde{\boldsymbol{z}}_{1:t})$
  Compute change-point probabilities: $p(r_t = 0, \tilde{\boldsymbol{z}}_{1:t})$
  Compute $p(\tilde{\boldsymbol{z}}_{1:t}) = \sum_{r_t} p(r_t, \tilde{\boldsymbol{z}}_{1:t})$
  Compute $p(r_t|\tilde{\boldsymbol{z}}_{1:t})$
  for $r_{t-1} = 1$ **to** $t + 1$ **do**
    Sample $\eta_t^s \sim p(\eta_t|\boldsymbol{\lambda}_t^{s-1}, r_{t-1}, \tilde{\boldsymbol{z}}_{1:t})$
    Sample $\boldsymbol{\lambda}_t^s \sim p(\boldsymbol{\lambda}_t|\eta_t^s, r_{t-1}, \tilde{\boldsymbol{z}}_{1:t})$ using the RW-MH algorithm
  end for
end for

by

$$p(\eta_t|r_{t-1}, \tilde{\boldsymbol{z}}_{1:t-1}, \boldsymbol{\lambda}_t^{s-1}) \propto p(\eta_t)p(\tilde{\boldsymbol{z}}_{1:t-1}|r_{t-1}, \eta_t^{s-1}, \boldsymbol{\lambda}_t^{s-1}) = \mathrm{Ga}(\eta_t|\kappa, \nu) \prod_{\tau=1}^{r_{t-1}} \mathrm{Dir}(\tilde{\boldsymbol{z}}_\tau|\eta_t^{s-1}, \boldsymbol{\lambda}_t^{s-1}), \tag{4.24}$$

and

$$p(\boldsymbol{\lambda}_t|r_{t-1}, \tilde{\boldsymbol{z}}_{1:t-1}, \eta_t^{s-1}) \propto p(\boldsymbol{\lambda}_t)p(\tilde{\boldsymbol{z}}_{1:t-1}|r_{t-1}, \eta_t^{s-1}, \boldsymbol{\lambda}_t^{s-1}) = \mathrm{Dir}(\boldsymbol{\lambda}_t|\boldsymbol{\beta}) \prod_{\tau=1}^{r_{t-1}} \mathrm{Dir}(\tilde{\boldsymbol{z}}_\tau|\eta_t^{s-1}, \boldsymbol{\lambda}_t^{s-1}). \tag{4.25}$$

where η_t^{s-1} and $\boldsymbol{\lambda}_t^{s-1}$ are the realizations drawn in the previous iteration of the Gibbs sampler. Moreover, since there is no direct way to obtain samples from the conditionals in Eq. (4.24) and Eq. (4.25), we propose to the use the Gibbs-within-Metropolis Hastings sampler presented in Martino et al. (2015, 2018). The main idea behind is to employ the random-walk (RW) version of the Metropolis Hastings (MH) algorithm (Martino and Elvira, 2017) to generate samples from Eq. (4.25), as this distribution typically becomes extremely narrow when the number of latent classes, K, is large. On the other hand, samples from the conditional distribution of Eq. (4.24) may be obtained in a simpler manner from the standard MH sampler with a Gamma *proposal*.

The complete CPD algorithm within the FPO inference model, including the Gibbs sampler and its versions, is summarized in Alg. 3. An important detail to consider within the present FPO approach is that given a huge number of latent classes K, the inference process could be still time-demanding as a consequence of approximating $\pi_t^{(r)}$. The particular issue could be the need of a large amount of samples S, at each time-step t.

4.2.3 Robustness to Missing Data

The PEO model also allows us to consider missing temporal data. Particularly, we assume that all missing observations in the sequence $\boldsymbol{x}_{1:t}$ are of type MCAR, that is, *missing completely at random*, which only denotes the lack of correlation among them. We denote each missing sample as $\boldsymbol{x}_t^m$. Moreover, samples that are fully observed are denoted as $\boldsymbol{x}_t^o$. The division of

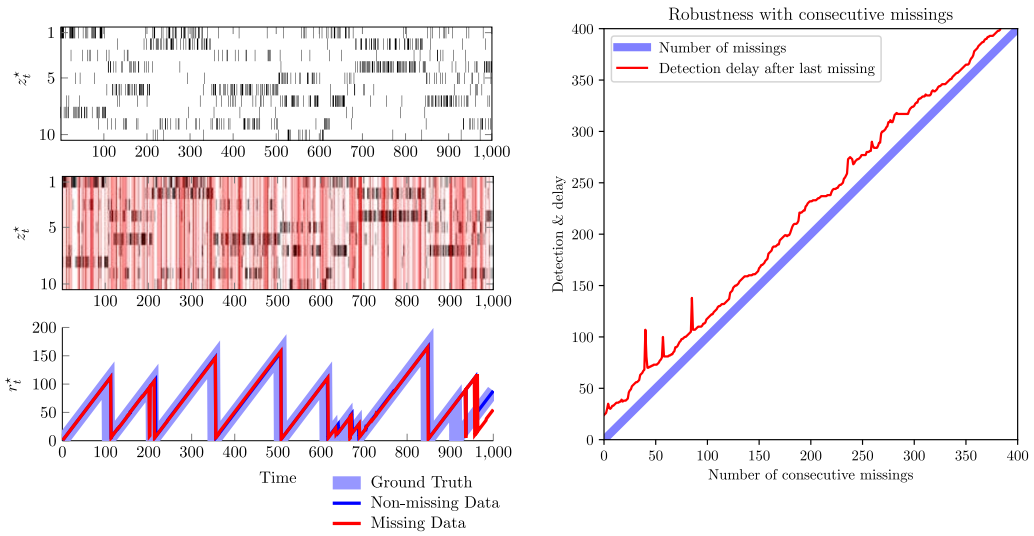


Figure 4.4: Change-point detection on the simplified hierarchical model with missing data. (LEFT TOP ROW) Fully observed sequence $\mathbf{z}^{\star}_{1:t}$ of point-estimates (1-of-K encoding) with $K = 10$. (LEFT MIDDLE ROW) Sequence $\mathbf{z}^{\star}_{1:t}$ with random missing entries (red) (25% rate). (LEFT BOTTOM ROW) True $r_{1:t}$ (grayed blue), MAP estimation of change points from the complete sequence (blue) and from the incomplete sequence (red). (RIGHT) Evolution of change-point detection delay as the number of consecutive missings increases.

the observational set into two subsets also translates the problem to the sequence of latent variables. In this case, we divide them into *lost* variables z^m_t, where all components of $\boldsymbol{x}_t$ are missing, and *observed* variables z^o_t, which correspond to the rest of cases.

Following a Bayesian approach, we marginalize the missing latent variables z^m_t out in the hierarchical CPD model. Then, the corresponding predictive probability is therefore reduced to $\Psi^{(r)}_t = \int p(z^m_t|r_{t-1}, \boldsymbol{z}_{1:t-1})dz^m_t = 1$. This still allows us to maintain the recursive methodology and hence, to compute the posterior distribution $p(r_t|\boldsymbol{x}_{1:t-1})$, even if a certain z_t is missing. Moreover, this recursivity remains unaltered since $p(r_t|r_{t-1})$ is always evaluated in a sequential manner. The corresponding expression for the joint distribution is as follows

$$p(r_t, \boldsymbol{z}_{1:t}) = \sum_{r_t - 1} p(r_t|r_{t-1})p(r_{t-1}, \boldsymbol{z}_{1:t-1}). \tag{4.26}$$

This approach for incomplete sequences presents good performance if the missing entries do not appear as long *chunks*. If that were the case, the time-step t for detection would *only* be delayed, as the influence of very old data in $p(r_t, \boldsymbol{z}_{1:t})$ in Eq. (4.26) decreases with the apparition of missing samples. This fact is illustrated in Fig. 4.4 where we compare the standard PEO version of the hierarchical model with and without missings. Additionally, we included an analysis of the CPD robustness to periods of consecutive missing observations. The results are also included in Fig. 4.4. Thus, we can confirm with this that the hierarchical method is robust under the presence of an arbitrary number of consecutive missing samples, with an approximately similar delay as in the *fully* observed case. If more missing data appeared, the delay would augment linearly in the same order as the increment of lost observations.

Algorithm 4 – Infinite-dimensional Hierarchical cpd

Input: Observe x_t and initialize $\hat{\boldsymbol{\varphi}}_{K_{t-1}}$.
Sample $z_t \sim p(z_t|\boldsymbol{z}^\star_{1:t-1})$
if $z_t = K_{t-1} + 1$ **then**
  Initialize $\hat{\boldsymbol{\varphi}}_{K_{t-1}+1}$
end if
Compute $p(z_t = k|\boldsymbol{z}^\star_{1:t-1}), \forall k \leq K_{t-1} + 1$
Compute $\mathbb{E}[\mathbb{I}\{z_t = k\}|\boldsymbol{z}^\star_{1:t-1}, x_t, \hat{\boldsymbol{\varphi}}_k^{(t-1)}], \forall k \leq K_{t-1} + 1$
Update parameters $\{\hat{\boldsymbol{\varphi}}_k\}_{k=1}^{K_{t-1}+1}$ using Eq. (4.31)
Calculate $z_t^\star = \arg\max(p(z_t|\boldsymbol{z}^\star_{1:t-1}, x_t, \{\hat{\boldsymbol{\varphi}}_k^{(t)}\}_{k=1}^{K_{t-1}+1})$
if $z_t^\star = K_{t-1} + 1$ **then**
  $K_t = K_{t-1} + 1$
end if
for $r_t = 1$ **to** t **do**
  Evaluate $\Psi_t^{(r)}$ using Eq. (4.29)
  Calculate $p(r_t, \boldsymbol{z}^\star_{1:t})$
  Obtain $p(\boldsymbol{z}^\star_{1:t}) = \sum_{r_t} p(r_t, \boldsymbol{z}^\star_{1:t})$
  Compute $p(r_t|\boldsymbol{z}^\star_{1:t})$
  Update $m_{k,t}^{(r)} \leftarrow m_{k,t-1}^{(r)} + \mathbb{I}\{z_t^\star = k\}$
end for
Return: $r_t^\star = \arg\max p(r_t|\boldsymbol{z}^\star_{1:t})$

4.2.4 Infinite Hierarchical Change-Point Detection

A general problem of the hierarchical cpd approach presented in the previous Sec. 4.2 is that the number of latent *classes*, K, must be known and fixed *a priori*. That is, K is not allowed to vary over time, which can be a stringent condition in several scenarios. For instance, in continual learning setups. In this section, we consider the more interesting case where K is unknown and can be time-varying, e.g. new latent classes z_t may appear as $t \to \infty$. Then, we cannot select the order of the latent model in advance.

A naive idea would be to fix an upper bound on K and proceed as in the previous described methodologies. However, this upper bound could not be available and, even if it is, the performance could be poor due to the high number of parameters $\boldsymbol{\theta}_t$ set from the very beginning. Therefore, the idea is to present a method for *unbounded* and time-varying K, that is, a maximum size K that is incremented only when unseen events are observed.

Using a potentially unbounded number of classes in the hierarchical cpd model results in the following problem when one integrates over $\boldsymbol{\theta}_t$ to compute $\Psi_t^{(r)}$ in Eq. (4.15). Assuming a Dirichlet distribution as the prior $p(\boldsymbol{\theta})$, which is the natural choice as conjugate-exponential system for categorical distributions, yields a tractable integral for the predictive posterior distribution. However, the evidence $p(z_t) \to 0$ as K grows. To overcome this issue, we can consider an exchangeable distribution of the form

$$p([\boldsymbol{z}_{1:t}]) = \sum_{\boldsymbol{z}_{1:t} \in [\boldsymbol{z}_{1:t}]} p(\boldsymbol{z}_{1:t}), \tag{4.27}$$

where $[\boldsymbol{z}_{1:t}]$ is a given division of classes, independent of the temporal assignments chosen. For instance, $\boldsymbol{z}_{1:3} = \{1, 2, 2\}$ would correspond to the same division of objects as $\boldsymbol{z}_{1:3} = \{2, 1, 1\}$. This is often known as the *exchangeability* property (Kingman, 1982; Pitman, 2002) and is a safe assumption in the hierarchical cpd framework as we are only interested in the probabilities of each z_t, rather than in the particular indexes of the sequence $\boldsymbol{z}_{1:t}$.

Chinese-Restaurant Process for CPD

The latent class model with an unbounded number of dimensions can be addressed via the *Chinese-restaurant process* (CRP) (Pitman, 2002), which is a *Bayesian nonparametrics* method (Orbanz and Teh, 2010). The CRP is based on the metaphor where clients (observations $\boldsymbol{x}_t$) are assigned to different tables (latent classes z_t) in a sequential manner. The assignment of classes to objects in the CRP is exclusively determined by the predictive posterior distribution, which is given by

$$p(z_t = k|z_1, \ldots, z_{t-1}) = \begin{cases} \frac{m_{k,t-1}}{t-1+\alpha}, & k \leq K_{t-1}, \\ \frac{\alpha}{t-1+\alpha}, & k = K_{t-1}+1, \end{cases} \tag{4.28}$$

where $m_{k,t-1}$ counts the number of assignments to each kth class up to time $t-1$. On the other hand, K_{t-1} is the number of classes associated with $m_{k,t-1} > 1$ and α is a hyperparameter, which corresponds to the natural parameterization of a symmetric Dirichlet prior distribution. It controls how likely is the appearance of a new class in the sequential model.

Exploiting the aforementioned CRP construction, one may still compute $\Psi_t^{(r)}$ in Eq. (4.14), and it is given by

$$\Psi_t^{(r)} = p(z_t^\star = k|r_{t-1}, \boldsymbol{z}_{1:t-1}^\star), \tag{4.29}$$

where we now count the number of MAP estimates, $z_t^\star$, equal to k up to time $t-1$. Notice that this expression is analogous to Eq. (4.28) for a given *run-length*, i.e. for each parallel thread in Fig. 4.1. Then, we may proceed to compute $p(r_t|\boldsymbol{z}_{1:t}^\star)$.

One key comment is that, so far, we have derived a tractable recursive method to introduce infinitely large latent class models into hierarchical CPD. However, nothing has been said here about the process of computing MAP estimates in a sequential fashion, required in Eq. (4.14). As the reader should note, this point has a connection with *continual learning*. This task is inspired in the previous Chap. 3 and presented in the next section.

Continual Learning of the CRP

The goal now is to compute MAP estimates of z_t in an online and recursive fashion. This task also involves the estimation of $\{\boldsymbol{\varphi}_k\}_{k=1}^{K_t}$, which are the parameters of the mapping between observations latent class variables, that is, $p(\boldsymbol{x}_t|z_t = k) = p(\boldsymbol{x}_t|z_t = k, \boldsymbol{\varphi}_k)$. Here, the number of classes K_t increases if when sampling from the CRP predictive equation, the results is $K_{t-1}+1$. That is, at the beginning of each iteration, we create a new class with emission probability given by Eq. (4.28), which is only kept if the MAP estimate indicates $z_t^* = K_{t-1}+1$. Since mixture models do not usually have closed-form solutions for the estimates of parameters, it is necessary to resort to the EM algorithm. Importantly, the prior distribution $p(\boldsymbol{z}_{1:t})$ factorizes according to

$$p(\boldsymbol{z}_{1:t}) = p(z_t|\boldsymbol{z}_{1:t-1})p(z_{t-1}|\boldsymbol{z}_{1:t-2}) \cdots p(z_1). \tag{4.30}$$

where we applied the chain-rule based on the contruction of the CRP. In addition, we have slightly modified the EM algorithm to accept the proposed continual learning framework. Concretely, the estimation of the parameter at each step is simply performed by taking one iterate of the *steepest* descent method, yielding a stochastic M-step (Cappé and Moulines, 2009). The E-step amounts to

$$\begin{aligned} \mathbb{E}[\mathbb{I}\{z_t = k\}|\boldsymbol{z}_{1:t-1}^\star, \boldsymbol{x}_t, \hat{\boldsymbol{\varphi}}_k^{(t-1)}] &= p(z_t = k|\boldsymbol{z}_{1:t-1}^\star, \boldsymbol{x}_t, \hat{\boldsymbol{\varphi}}_k^{(t-1)}) \\ &\propto p(\boldsymbol{x}_t|z_t = k, \hat{\boldsymbol{\varphi}}_k^{(t-1)}) p(z_t = k|\boldsymbol{z}_{1:t-1}^\star), \end{aligned}$$

where $\hat{\boldsymbol{\varphi}}_k^{(t)}$ is the estimate of $\boldsymbol{\varphi}_k$ at time t, and we have also exploited Eq. (4.28). In the M-step, the estimate of the parameters $\{\boldsymbol{\varphi}_k\}_{k=1}^{K_t}$ is updated based on the gradient:

$$\hat{\boldsymbol{\varphi}}_k^{(t)} \leftarrow \hat{\boldsymbol{\varphi}}_k^{(t-1)} + \eta_{k,t} \nabla_{\boldsymbol{\varphi}_k} \mathbb{E}[\mathcal{L}_{\boldsymbol{\varphi}}(\boldsymbol{x}_{1:t}, \boldsymbol{z}_{1:t})], \tag{4.31}$$

where $\eta_{k,t}$ is the (adaptive) learning rate for the kth class at time t. In this expression, we have assumed that the same initial learning rate is chosen for the parameters of a given class, but it is possible to select multiple learning rates per class. Once we have the E- and M-steps, we can compute the posterior of z_t and maximize it to obtain $z_t^\star$ as in Sec. 4.2. Finally, in Alg. 4 we present all the necessary computations of the proposed recursive method at each time instant t.

4.2.5 Robust Hierarchical Change-Point Detection

The full derivation of the joint distribution $p(r_t, \boldsymbol{z}_{1:t})$, given the hierarchical CPD model is explained in the previous Sec. 4.2, and its particular case for a potentially infinite number of classes K is developed in Sec. 4.2.4. However, working with sequences of MAP point-estimates $z_t^\star$ in both versions of the hierarchical CPD model might be sometimes problematic. It could lead to false-alarm or missing detection problems when the underlying inference process and particularly, the posterior distribution $p(\boldsymbol{z}_{1:t}|\boldsymbol{x}_{1:t})$ is extremely *flat*. That is, it presents more uncertainty than expected. One may notice that if MAP estimates do not coincide with the *true* assignments, then the CPD method will be much more noisy, leading to undesired results.

In this section, we describe a robust extension of the hierarchical CPD model to address a better characterization of the underlying distribution at each time-step t, when it is not well-fitted. For this task, we can generate *pseudo-observations* of the latent variables by drawing S i.i.d. samples of the posterior distribution as $z_t^{(1)}, z_t^{(2)}, \cdots, z_t^{(S)} \sim p(z_t|\boldsymbol{x}_t)\ \forall t$, rather than working with a single point-estimate $z_t^\star$ as we do in Sec. 4.2.

The new approach addresses the question of how to deal with a subset of S samples instead of just one (the point-estimate $z_t^\star$) at each time-step. A potential idea would be to consider Monte-Carlo (MC) approximations, but it would imply to draw $S \times t$ samples at each iteration. In the long term, this would be unfeasible for an efficient CPD method. Alternatively, we propose to draw samples from a multinomial distribution, which in our case preserves the original prior-conjugacy of Adams and MacKay (2007) while at the same time is consistent with the new hierarchical version in Moreno-Muñoz et al. (2018).

Multinomial Sampling for Posterior Characterization

A multinomial distribution with parameters $\boldsymbol{\theta}_t \in \mathcal{S}^K$ and N, measures the probability that each class $k \in \{1, \ldots, K\}$ is observed n_k times over N categorical independent realizations with success probabilities $\boldsymbol{\theta}_t$. This model allows us to perform an *augmentation* of the pseudo-observation set, with the cost of introducing one single parameter in the model: $N = S$, the total number of samples drawn from the posterior distribution $p(z_t|\boldsymbol{x}_t)$.

Given the sampled vector $\boldsymbol{z}_t^\star = [z_t^{(1)}, z_t^{(2)}, \ldots, z_t^{(S)}]^\top$, with $\boldsymbol{z}_t^\star \in \{1, 2, \ldots, K\}^S$, we can define its associated counting vector as $\boldsymbol{c}_t \in \mathbb{Z}_+^K$ where we also add $c_t^k := \sum_{s=1}^S \mathbb{I}\{z_t^{(s)} = k\}\ \forall k$. Thus, we have that $\sum_{k=1}^K c_t^k = S$. Having the subset of samples at each time-step t, we can see the vector $\boldsymbol{c}_t$ as an i.i.d. observation of a multinomial likelihood distribution. In particular, it would be parameterized now by the natural parameters $\boldsymbol{\theta}_t \in \mathcal{S}^K$ that are of our interest, and also with a parameter $S \in \mathbb{N}$ that we already know.

With the previous notation in the hand, we assume the following generative model,

$$\boldsymbol{\theta}_t \sim \text{Dirichlet}(\boldsymbol{\alpha}), \tag{4.32}$$

$$\boldsymbol{c}_t \sim \text{Multinomial}(\boldsymbol{\theta}_t, S), \tag{4.33}$$

where $\boldsymbol{\alpha} \in \mathbb{R}_+^K$ and the likelihood function for $\boldsymbol{c}_t$ is

$$p(\boldsymbol{c}_t|\boldsymbol{\theta}, S) = p(c_t^1, c_t^2, \ldots, c_t^K|\boldsymbol{\theta}_t, S) = \frac{S!}{\prod_{k=1}^K c_t^k!} \prod_{k=1}^K \theta_k^{c_t^k}. \tag{4.34}$$

The update rule of the posterior distribution parameters has the following closed-form $\boldsymbol{\alpha}' = \boldsymbol{\alpha} + \boldsymbol{c}_t$. This allows us to directly update them every time a new sample $z_t^{(s)}$ is observed.

Notice from the l.h.s. term in Eq. (4.34) and the formal definition of $\boldsymbol{c}_t$ that we are not working with probabilities over the S-dimensional sampled vectors, but over their permuted classes instead. That is, two sampled vectors are equivalent $\boldsymbol{z}_{S_1}^\star \sim \boldsymbol{z}_{S_2}^\star$, iff their associated counting vectors satisfy $\boldsymbol{c}_{S_1} = \boldsymbol{c}_{S_2}$. Thus, the vector $\boldsymbol{z}_{S_2}^\star$ must be a permutation of the vector $\boldsymbol{z}_{S_1}^\star$.

For performing hierarchical CPD, we now wish to infer the parameters $\boldsymbol{\theta}_t$ conditioned to each *run-length* hypothesis r_t and its associated data $\boldsymbol{x}_{1:t}$. To carry out with the inference methodology depicted in Fig. 4.1, we need to find the predictive posterior distribution for each r_{t-1} variable, that is, partition hypothesis. A first step is to marginalize out the natural parameters $\boldsymbol{\theta}_t$, then we have

$$p(\boldsymbol{c}_t|r_{t-1}, \boldsymbol{c}_{1:t-1}^{(r)}) = \int p(\boldsymbol{c}_t|\boldsymbol{\theta}) p(\boldsymbol{\theta}_t|r_{t-1}, \boldsymbol{c}_{1:t-1}^{(r)}) d\boldsymbol{\theta}_t, \tag{4.35}$$

where the predictive term is now $\Psi_t^{(r)} := p(\boldsymbol{c}_t|r_{t-1}, \boldsymbol{c}_{1:t-1}^{(r)})$, and has not closed-form solution. However, it is a function of the statistics of the model and its computation is straightforward,

$$\Psi_t^{(r)} = \frac{\Gamma(S+1)\Gamma_\alpha}{\prod_{k=1}^K \Gamma(c_t^k + 1)} \frac{\prod_{k=1}^K \Gamma(c_t^k + \alpha_{t-1}^k)}{\prod_{k=1}^K \Gamma(\alpha_{t-1}^k)\Gamma(S + S_\alpha)}, \tag{4.36}$$

where we have previously defined $S_\alpha := \sum_{k=1}^K \alpha_{t-1}^k$. Additionally, using both the binomial coefficient definition and the Gamma function property $\Gamma(n+1) = n!$ for $n \in \mathbb{N}$, we transform the previous expression in the following one:

$$\Psi_t^{(r)} = \binom{S + S_\alpha - 1}{S}^{-1} \prod_{k=1}^K \binom{c_t^k + \alpha_{t-1}^k - 1}{c_t^k}. \tag{4.37}$$

The term S_α grows linearly with S per time-step, leading sometimes to numerical instabilities in the l.h.s. term of Eq. (4.37) for large values of t. Therefore, we consider the following alternative, that is numerically more stable and is a result of the manipulations on the terms of Eq. (4.37), it is

$$\Psi_t^{(r)} = \prod_{k=1}^K \prod_{j=0}^{c_t^k - 1} \frac{\alpha_{t-1}^k + j}{S_\alpha + S_c^{(k-1)} + j} \frac{S_c^{(k-1)} + j + 1}{j+1}, \tag{4.38}$$

with $S_c^{(k-1)} := \sum_{k'=1}^{k-1} c_t^{k'}$ for $k = 1, 2, \ldots, K$. Importantly, notice from the previous expression and the general hierarchial CPD model of Eq. (4.14) that the computational cost always

Algorithm 5 – Robust Hierarchical CPD

Input: Observe $\boldsymbol{x}_t \rightarrow$ obtain $p(z_t|\boldsymbol{\psi}, \boldsymbol{x}_t)$
Sample $z_t^{(1)}, z_t^{(2)}, \ldots, z_t^{(S)} \sim p(z_t|\boldsymbol{\psi}, \boldsymbol{x}_t)$
Count and build $\boldsymbol{c}_t$
for $r_t = 1$ **to** t **do**
    Evaluate $\Psi_t^{(r)}$ using (4.37)
    Calculate $p(r_t, \boldsymbol{c}_{1:t})$
    Obtain $p(\boldsymbol{c}_{1:t}) = \sum_{r_t} p(r_t, \mathbf{c}_{1:t})$
    Compute $p(r_t|\boldsymbol{c}_{1:t})$
    Update $\alpha_{t+1}^k = \alpha_t^k + \boldsymbol{c}_t^k \quad \forall k \in \{1, ..., K\}$
end for
Return: $r_t^\star = \arg\max p(r_t|\boldsymbol{c}_{1:t})$

grows linearly with S. This is the main advantage of the method, a better characterization of the posterior $p(z_t|\boldsymbol{x}_t)$ with no extra complexity. Finally, in Alg. 5, we present all steps that must be followed to obtain the *run-length* estimates $r_t^\star$ from the initial sequence of high-dimensional observations $\boldsymbol{x}_{1:t}$.

The variable $r_t^\star$ corresponds to MAP point-estimates of the *run-length* at each-time step t. It is the variable that we will later use to show the performance of the CPD model, and it recursively represents the most likely CP in the sequence.

4.3 Evaluation of Hierarchical CPD Models

In this section, we evaluate the performance of the hierarchical CPD model (Sec. 4.2) and its adjacent versions presented in this chapter. In particular, they are both the infinite-dimensional adaptation of Sec. 4.2.4 and the robust hierarchical CPD model of Sec. 4.2.5. For the evaluation in this chapter, we will focus on synthetically generated data that we use for proving the performance of the models. The empirical validation on real-world data, i.e. with application to human behavior learning and behavioral change detection, is included in Chap. 5.

4.3.1 Hierarchical CPD Simulation

In the first experiment of this experimental section, we validate the hierarchical CPD method using synthetic data. Particularly, we generate discrete sequences of *known* latent variables $\boldsymbol{z}_{1:t}$ with $T = 500$ instances and a dimension of $K = 5$ classes (future z_t true assignments). In this sequence, we introduce four change-points. Then, the generative parameters $\boldsymbol{\pi}_t$ for each partition are generated by sampling $\boldsymbol{\pi}_t \sim \text{Dir}(\alpha/5, \ldots, \alpha/5)$, where we fixed $\alpha = 25$. The resulting sequence $\boldsymbol{z}_{1:t}$ is used to generate pairs of multivariate Bernoulli-Gaussian observations following the heterogeneous latent-class model described in Chap. 2.

In this particular experiment, we only observe the given stream of binary and real-valued observations, never the *true* sequence of latent variables $\boldsymbol{z}_{1:t}$ that we now aim to infer. To obtain the posterior probabilities over the class assignments, $p(z_t|\boldsymbol{x}_t)$, we apply the EM algorithm described in the previous Chap. 2 and Chap. 3. Importantly, as we use here the circadian covariance function model (Sec. 2.1.4), when learning from samples, we truncated the number of Fourier coefficients to $C = 3$, while the data was initially generated with $C = 2$. Later, from the sequence of posterior probability vectors, $p(\boldsymbol{z}_{1:t}|\boldsymbol{x}_{1:t})$, point-estimate pseudo-observations were taken using the *map* estimator of Sec. 4.2.1. We run the whole process for multiple initializations and K classes in the Gaussian-Bernoulli latent class model.

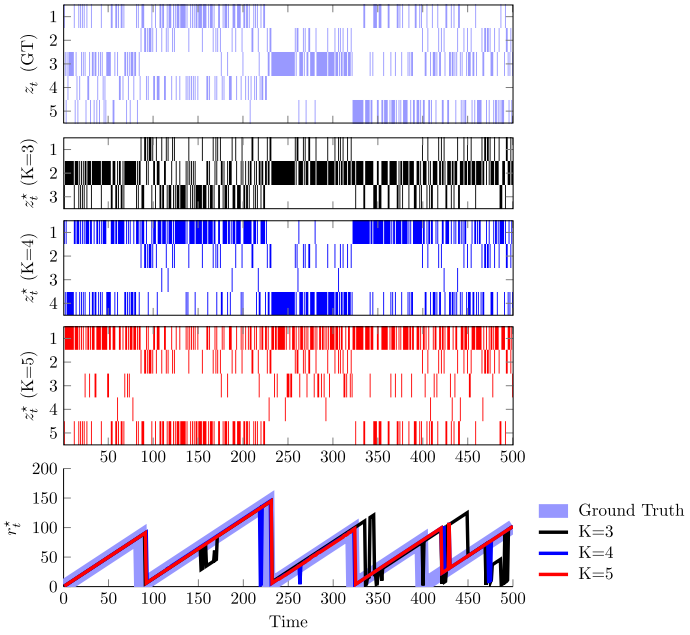


Figure 4.5: Results for the synthetic experiment. Each observation $\boldsymbol{x}_t$ modelled by z_t was heterogeneous, i.e. real-valued and binary features, with a dimensionality $D = 24$. (FIRST ROW) Sequence of true latent variables. (SECOND TO FOURTH ROWS) Sequence of MAP estimates of the latent variables for $K = 3, 4$, and 5. (LAST ROW) Ground truth of CP segments and MAP-CPD traces of $r_t^\star$ for $K = 3, 4$, and 5 latent classes.

ORDER	$\mathcal{Q}$/LOG-LIK. (av. $\pm$ std.)	BIC	AIC
$K = 2$	$-20.87 \times 10^3 \pm 23.83$	-21.11×10^3	-20.97×10^3
$K = 3$	$-20.36 \times 10^3 \pm 158.45$	-21.11×10^3	-20.92×10^3
$K = 4$	$-20.35 \times 10^3 \pm 146.30$	-21.10×10^3	-20.88×10^3
$K = 5$	$-20.26 \times 10^3 \pm 56.41$	$\underline{-20.77 \times 10^3}$	$\underline{-20.46 \times 10^3}$
$K = 6$	$-20.27 \times 10^3 \pm 27.22$	-20.86×10^3	-20.49×10^3
$K = 7$	$-20.24 \times 10^3 \pm 62.18$	-20.99×10^3	-20.56×10^3

Table 4.1: Average log-likelihood, and largest BIC and AIC metrics for 10 initializations. Best BIC and AIC models are underlined.

In Fig. 4.5, we show the resulting estimates $z_t^\star$ as well as $r_t^\star$ for $K = \{3, 4, 5\}$ and the *ground-truth* of partitions. The proposed hierarchical method is able to detect with high accuracy, although, as the number of categories K decreases, the detection presents a longer delay. Our hypothesis is that, as long as the circadian model gets a sufficiently large number of latent classes (e.g. it becomes more flexible and representative), the CPD model better identifies adjacent partitions. Moreover, it is possible to choose the model order, K, in the hierarchical CPD method using the Bayesian and *Akaike* information criterion (BIC, AIC). In Tab. 4.1, we show the average and standard deviation per order, which in this case, it indicates the correct dimensionality of the discrete latent space.

Comparison with CPD Alternatives for High-Dimensional Data

This section shortly evaluates and compares the performance of the proposed method with alternative CPD techniques in the literature. Particularly, we consider: i) the Bayesian online change-point detection (BOCPD) algorithm, ii) principal component analysis (PCA) preprocessing followed by BOCPD (PCA-CPD), iii) optimal partitioning (OP) (Jackson et al., 2005) and iv) binary segmentation (BS) (Scott and Knott, 1974). Since the BOCPD and the PCA-CPD methods work directly on the observational set $\boldsymbol{x}_{1:t}$, in the comparison with proposed detector, we have considered only Gaussian observations. On the contrary, the OP and BS methods work directly on $\boldsymbol{z}^{\star}_{1:t}$, and we can therefore use the Gaussian-Bernoulli model of Sec. 2.1.2. Moreover, for the BOCPD algorithm, we have assumed a diagonal covariance model, since a full-covariance matrix would present extremely poor detection performance as the dimensionality increases and CPD cannot obtain enough evidence. Precisely this problem is what we try to avoid by combining the BOCPD with a dimensionality-reduction step based on PCA as in the second case. We keep the two components with largest variance, which yields a transformed observation sample with only two dimensions. Then, we use the diagonal covariance model within the BOCPD algorithm. The spirit of this dimensionality reduction idea is similar to the one proposed in the hierarchical CPD framework. However, it considers a continuous low-dimensional manifold with an orthogonal linear projection instead.

We evaluate the performance of each one of the CPD methods and the resulting metrics are included in Tab. 4.2. Particularly, results are given for Gaussian observations of dimension D, (with acronym $\mathcal{N}_D$) and the Gaussian-Bernoulli model. Here, each kind of heterogeneous observation is also of dimension D (with acronym $\mathcal{NB}_D$). Then, in the latter case, the data samples have a total dimensionality of $2D$. Additionally, we considered two metrics, delay and CP detection rate, for which we assumed that a CP is detected iff the method identifies the CP location at the time-step it appears or afterwards. The data in all these runs were synthetically generated as in the previous experiment. We set five CPs every $t = 100$ steps.

The results in Tab. 4.2 show that the proposed hierarchical CPD method presents higher detection rate than the competitors with sufficiently small delay. Moreover, its main strengths compared to the original model of Adams and MacKay (2007), is that it allows to consider extremely high-dimensional data, with heterogeneous features with arbitrary statistical data-types and even circadian structure without decreasing the detection performance.

4.3.2 Infinite Hierarchical CPD Simulation

In this section we evaluate the performance of the proposed method. We apply the infinite-dimensional hierarchical CPD algorithm to real-world data, and in particular, to a sequence of raw nuclear magnetic response measurements taken during a well-drilling process. These data have been previously used in the context of time-series modelling and CPD analysis by Adams and MacKay (2007); Fearnhead and Clifford (2003); JK and WJ (1996). The data consists of 4500 real-valued univariate observations taken at a fixed sampling frequency. In the following, we assume that the time-steps are discrete and ordered for simplicity.

To apply the proposed model, we first choose $p(\boldsymbol{x}_t|z_t = k, \boldsymbol{\varphi}_k)$ to be Gaussian distributed with unknown mean and variance, that is, $\boldsymbol{\varphi}_k = \{\mu_k, \sigma_k^2\}$. Moreover, the model has two extra hyperparameters that we need to select. The first one, which is related to the CPD method, is the parameter λ of the hazard function that is used as the conditional prior, $p(r_t|r_{t-1})$. In the experiments, we have selected $\lambda = 10^6$. The second one is the parameter α, which is involved in the CRP construction, and controls how likely is the appearance of a new unseen class. We set it to $\alpha = 1.0$. For the stochastic M-step, we use two different adaptive learning rates for the mean and variance whose initial values are given by $\eta_\mu = 1.0$ and $\eta_\sigma = 0.02$.

	BOCPD		PCA-CPD		HIERCPD	
model	delay	rate	delay	rate	delay	rate
$\mathcal{N}_2$	22.30 ± 20.02	0.92	24.35 ± 25.51	0.68	22.31 ± 18.79	0.88
$\mathcal{N}_5$	18.41 ± 9.88	0.96	20.42 ± 13.57	0.76	12.56 ± 6.51	1.0
$\mathcal{N}_{10}$	10.54 ± 8.85	0.96	14.45 ± 12.21	0.8	8.71 ± 4.83	0.96
$\mathcal{N}_{20}$	7.42 ± 7.61	0.84	15.45 ± 10.22	0.88	8.76 ± 5.35	1.0
$\mathcal{N}_{50}$	5.05 ± 2.97	0.76	11.81 ± 10.49	0.8	8.39 ± 3.34	1.0
$\mathcal{N}_{100}$	6.36 ± 1.93	0.76	13.75 ± 13.47	0.8	9.56 ± 7.10	1.0
$\mathcal{N}_{200}$	4.41 ± 1.18	0.6	12.06 ± 9.45	0.64	13.39 ± 12.27	1.0

	OP		BS		HIERCPD	
model	delay	rate	delay	rate	delay	rate
$\mathcal{NB}_2$	4.26 ± 6.06	0.6	1.25 ± 0.95	0.16	25.75 ± 21.26	0.8
$\mathcal{NB}_5$	2.07 ± 2.32	0.56	1.02 ± 1.22	0.2	15.19 ± 9.98	0.84
$\mathcal{NB}_{10}$	0.83 ± 1.21	0.72	2.01 ± 1.82	0.16	13.86 ± 7.32	0.92
$\mathcal{NB}_{20}$	1.06 ± 1.38	0.64	2.50 ± 1.73	0.16	13.08 ± 10.35	1.00
$\mathcal{NB}_{50}$	1.01 ± 0.97	0.68	1.01 ± 1.58	0.36	11.21 ± 6.67	1.00
$\mathcal{NB}_{100}$	1.91 ± 3.26	0.84	0.77 ± 0.83	0.36	13.83 ± 13.58	0.96
$\mathcal{NB}_{200}$	1.61 ± 2.06	0.84	3.04 ± 5.81	0.36	13.54 ± 8.91	0.96

Table 4.2: Hierarchical CPD (HIERCPD) vs. benchmark of change-point detection methods: Bayesian change-point detection (BOCPD), principal component analysis (PCA-CPD), optimal partitioning (OP) and binary segmentation (BS). The upper table shows the results for Gaussian observations with different dimensionality ($\mathcal{N}_D$), and the lower table shows the case with heterogeneous data Gaussian-Bernoulli ($\mathcal{NB}_D$). Best delay and rate metrics are underlined.

Importantly, we made both learning rates decrease at a rate of 2% per time-step if $z_t = k$ was selected as the most likely latent class. This choice avoids adapting very old parameters with new incoming data.

Fig. 4.6 shows the results obtained for $t = 4500$ iterations.[1] The unbounded model is compared with the hierarchical CPD approach with an upper bound on the number of classes $K = 10$. In the upper figures in Fig. 4.6 we can see the well-drilling signals, as well as the latent-class assignments in different colors for both approaches. As can be observed, the final number of classes inferred by the CRP was $K_{4500} = 7$. In the bottom figures we show the MAP estimates of the run-length, $r_t^\star$. These figures show that the MAP estimation of the run-length is well aligned with the signal transitions. Furthermore, it should be noted that the proposed method is more robust to outliers as can be seen for $t \approx 200$ and $t \approx 600$, where the outlier is captured by the latent class assignment but a CP is not declared. In fact, the MAP estimate of the run-length is noisier for the method with a fixed number of classes than for the unbounded model.

In addition, the latent-class assignments look more consistent in the case of the infinite-dimensional hierarchical CPD algorithm, where both the initial and final samples of the well-drilling signal are assigned to the same latent-class. The main two advantages of the method

[1] A video demonstration of the complete simulation of the algorithms is available at https://www.youtube.com/watch?v=ymZPNURhtIc.

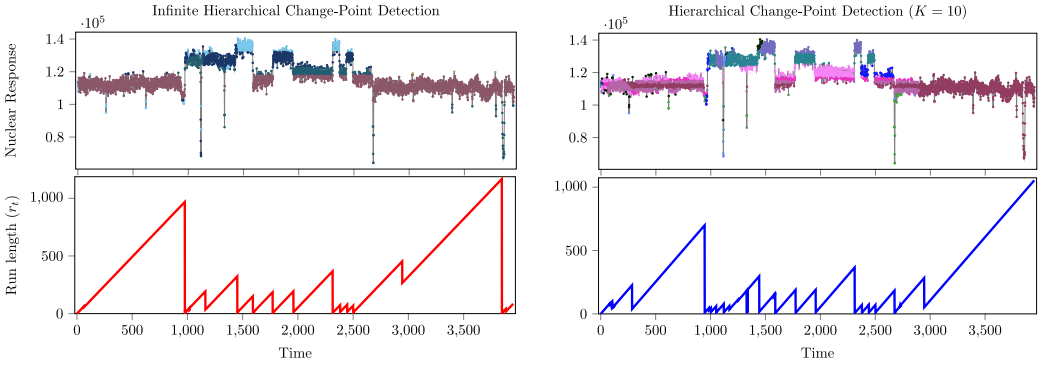


Figure 4.6: Upper row plots show the well-drilling univariate signal for the unbounded latent variable model (left) and the hierarchical CPD method (right) with fixed K. The colours represent latent-class asignments. Bottom row plots show the MAP estimates of the run-length.

can be observed from the empirical results. First, the method uses past learned parameters to infer assignments over very recent data, that is, assignments coincide along time. Second, the CRP is able to discover new unseen latent-classes without fixing the model complexity *a priori* and avoids the overlapping with previous discovered classes. For instance, if a new latent class k^* appears, it would not coincide with the previous learned ones, and neither their parameters.

4.3.3 Robust Hierarchical CPD Simulation

In this section, we evaluate the performance of the proposed multinomial sampling extension for hierarchical CPD. Particularly, we study the improvements of the method (named here Multinomial CPD), over synthetically generated data. The core idea of the experiment is that we may increase or decrease the quality of inference artificially over the latent variables to prove that detection is still reliable. In all the experimental results, we consider that a CP is detected at a time-step $t = t'$ iff there is an abrupt decrease from $r^\star_{t'-1}$ to $r^\star_{t'}$, which means that the CP ocurred at instant $t = t' - r^\star_{t'}$. We set $r^\star_t < r^\star_{t-1}$ as the condition for detection.

The Multinomial CPD model has been applied to sequences of synthetic data and the results have been summarized in Fig. 4.7 and Tab. 4.3. In the these cases, we evaluated the performance of the method for several sampling sizes S, drawn at each time-step and for different levels of *flatness* given the generative posterior distribution.

We have also fixed the number of CPs on the latent sequence to five, that is, six partitions. Each one occurring every 100 time-steps. Moreover, we have run the algorithm for a total period of $T = 600$. In this experiment, the posterior distribution $p(z_t|\boldsymbol{x}_t)$ over the latent variables is simulated. For each partition ρ, we generated a set of 100 K-dimensional vectors $\boldsymbol{\theta}_\rho$, from a Dirichlet distribution with natural parameters $\boldsymbol{\beta}_\rho$. At the same time, these 6 subsets of parameters $\boldsymbol{\theta}_\rho$ were sampled from an uniform distribution in the interval $(0, \eta)$. Here, the hyperparameter η defines the *flatness* of the synthetic posterior distribution. Intuitively, a lower η value would imply a flatter generative distribution of the simulated posterior probabilities. The hyperparameter K was fixed to 20 classes for the whole experiment. In the proposed model, each S-dimensional vector is sampled from a distribution $\text{Mult}(\boldsymbol{\theta}_\rho, S)$, where $\boldsymbol{\theta}_\rho$ is the vector previously presented.

The conditional prior probability of the run-length r_t is a function of the hyperparameter

η	$S = 10$ CPD RATE	$S = 50$ CPD RATE	$S = 100$ CPD RATE	HIER. CPD RATE
2.0	-	0.12	<u>0.32</u>	-
3.0	0.52	<u>0.88</u>	0.84	0.2
4.0	0.88	0.96	<u>1.0</u>	0.76
10.0	0.96	1.0	<u>1.0</u>	0.96

η	$S = 10$ DELAY	$S = 50$ DELAY	$S = 100$ DELAY	HIER. DELAY
2.0	∞	5.33 ± 2.30	<u>5.37 ± 1.59</u>	∞
3.0	5.30 ± 2.09	5.68 ± 3.01	<u>4.20 ± 2.17</u>	10.0 ± 7.87
4.0	3.57 ± 2.15	3.28 ± 2.53	<u>2.30 ± 0.96</u>	$5.27 + 2.00$
10.0	2.06 ± 1.77	1.32 ± 0.39	<u>1.31 ± 0.40</u>	3.52 ± 2.00

Table 4.3: Multinomial CPD vs. Hierarchical CPD metrics. All delay values ($\times 10$). Best delay and rate metrics are <u>underlined</u>.

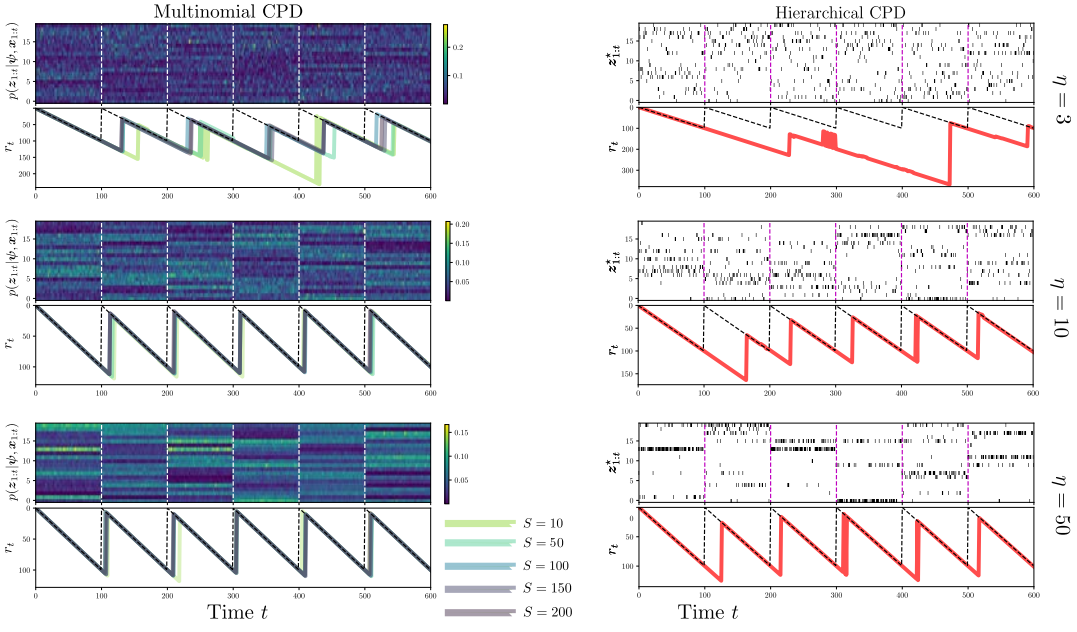


Figure 4.7: Comparison between the multinomial CPD, based on sampling from the latent class posterior, and the baseline hierarchical CPD method. The resulting CPs (bottom figures) are considered as jumps over the MAP estimates (solid lines) of the run-length $r_t \, \forall t$. Dashed lines indicate the true change points. (LEFT COLUMN) Each row represents an example with a more or less flat posterior distribution (upper figures) indicated by η. Colors of the r_t lines indicate the number of samples S used. (RIGHT COLUMN) Results for CPD from different point-estimate pseudo-observations $\boldsymbol{z}^{\star}_{1:t}$ (upper figures).

λ^{-1}, which controls the initial probability of a change. Thus, the higher λ is, the less probable a change is. For this experiment and having the Multinomial CPD model, we have defined it

as a function of the number of samples $\lambda = 10^S$ to make both comparable the terms involved in Eq. (4.14) and also the results in the experiment for different number of samples. The intuition behind this choice is that, for high values of S, we want the prior probability of a CP to be almost zero. So that the CP occurrence is mainly determined in a data-driven manner. However, more accurate results could be found if we tune the hyperparameter λ at each particular case. For the comparison in Tab. 4.3 with the hierarchical CPD version, we have consired the same setupt except for λ. It has been fixed to 10^{20} for stability purposes.

In Fig. 4.7, we compare the Multinomial CPD approach (left column) for different number of samples $S = \{10, 50, 100, 150, 200\}$ with the hierarchical CPD method (right column). Additionally, we tried several levels of *flatness* in the simulated posterior distribution $p(z_t|\boldsymbol{x}_t)$ via $\eta = \{3.0, 10.0, 50.0\}$ (rows). In the upper figures, we show the posterior distributions of the latent variables z_t or their MAP assignments in the left colum case. In the lower figures, the MAP point-estimates of the *run-length* r_t are jointly shown within dashed lines indicating the true location of CPs.

We have also summarized the delay and detection metrics of running the method under several initializations in Tab. 4.3. Each result is given for pair of values (S, η). In particular, we show the average metrics of *precision*, defined as the ratio of CPs detected for each pair (S', η'), and the mean and std. deviation of the *delay*. This is defined as the number of time-steps between the detection event $t = t'$ and the CP location $t = t' - r^\star_{t'}$ indicated. For instance, if a CP is detected at $t = 150$ and $r^\star_{150} = 30$, it would mean that the CP occurred at $t = 120$, and the delay would be of 30 steps.

4.4 Discussion

In this chapter, we developed a novel generalization of the well-know Bayesian change-point detection method of Adams and MacKay (2007), the BOCPD algorithm. Particularly, the new framework handles heterogeneous high-dimensional observations with any arbitrary combination of likelihood functions as well as an unknown periodic structure. The set of proposed techniques makes two main contributions to the state-of-the-art. The first one, that we denote as the *hierarchical detector*, is the probabilistic extension of the Bayesian CPD method to accept latent variable models. The main benefit of this strategy is that the CPD method now works with less complex discrete data, as the CPs are easier to be detected in the latent representation. In the second case, our *infinite* hierarchical method prevents the limitation of fixing the order of the latent variables. This is, we introduced the CRP to allow for an unbounded number of classes. In the context of human behavior learning, this is a key point of contribution as we will discuss later in Ch. 5. Finally, we introduced a multinomial sampling method that improves the detection rate and reduces the delay when using the hierarchical CPD. The experimental results show significant improvements in the detection as posterior estimates become less certain if high-dimensional data is difficult characterize. Interestingly, even under good performance, the multinomial sampling method reduces the detection delay, what in practice is a key point for its application to human behavior in medicine.

CHAPTER 5

BEHAVIOR CHANGE DETECTION

THE fundamental idea behind the application of probabilistic modelling to human behavior characterization is the problem of *change* in medicine. As we previously explored in Ch. 4, once a model is accurately learned in a data-driven manner, one can also detect *anomalies*, outliers or changes given its adjacent fitted parameters. This way of determining whether or not a point of variability exists is of particular importance in the medical applications that we may consider in this thesis. Thus, the application of the technical contributions developed in the previous chapters (Ch. 2, Ch. 3 and Ch. 4), and particularly change-point detection (CPD), is directly oriented to mental health applications in this chapter and particularly, to the *passive* detection of events.

Here, we focus on psychiatric patients, in those clinical cases with severe affective disorders with higher prevalence, i.e. depression or bipolar diseases. As long as this sort of diseases become chronic, the life conditions of sufferers show patterns of perpetual disabilities in social, labour and residential domains. This problem even worsens as they can be unaware of their own disability or the symptoms that are harbingers on an imminent crisis. Our purpose in this chapter is to provide an *unobtrusive* early detection methodology for relapses, which is based on both statistical modelling and change-point detection. The final goals can be summarized in: i) building intepretable representations of daily patients behavior for a better comprehension and assisstance to clinicians, ii) detection of critical events, e.g. change-points in the sequence of low-dimensional representations and iii) validation of the detected events within clinical data from interventions in hospitals and urgencies.

We begin in Sec. 5.1 with the presentation of the first approach. We identify the mobility patterns given a first collection of five data records. Additionally, in Sec. 5.1.1, we describe the development of the EB2 study and the monitoring system for patients, which we will later use for obtaining the indicators of the behavioral and clinical changes in a tool-based manner. In the next Sec. 5.2, we extend the previous results to a wider range of patients as well as we accept less pre-processed data. The final outcome is a unobtrusive system that deals with both heterogeneous statistical data types and missing values. We also obtain early indicators of crisis events from raw monitoring traces. These were recorded from mental health patients with *at least* one previous suicide attempt registered. The previous estimates are validated in Sec. 5.3, where we compare the detected change-points with the *true* events registered by clinicians in their hospitalary systems.

The main core of the work in this chapter have been previously presented as *peer-review* papers in both venues and journals. The first one, Berrouiguet et al. (2018), was published in the Journal of Medical Internet Research (JMIR), in the section of MHealth and UHealth. This work was also a collaboration with D. Ramírez and several clinicians from mental health departments, both national and international. Particularly, we remark the collaboration with the team led by Dr. E. Baca-García and Dr. Sofian Berrouiguet from the CHRU C. V. University Hospital of Brest. The second piece of work, Moreno-Muñoz et al. (2020c), was recently presented in the Machine Learning for Mobile Health (ML4MH) Workshop at the international conference NeurIPS in December 2020. This was also a collaboration with several post-graduate colleagues, including A. Moreno, L. Romero-Medrano and J. Herrera.

5.1 Behavior Change Detection: A First View

The recent development of mobile electronic devices such as personal smartphones or wearables has gained important attention in healthcare applications due to their ubiquitous conditions for *pervasive* sensing. It is currently known as electronic health (e-Health) in the literature. Particularly, the disruption of smartphones afforded new opportunities (Miller, 2012) to obtain objective, reliable and real-time monitoring data of patients outside the ambulatory domain where standard assessment methodologies cannot be driven in a daily manner.

However, the use of mobile electronic devices for medical studies, and particularly behavior-oriented, is not new. For about 20 years, researchers have been performing advances on this side. To name a few, some type of studies in collaboration with telecommunications companies gathered aggregated location traces, meta-data of call records, dial numbers and even their duration. This valuable information helped to model the social interactions of users as well as their routines of mobility (Calabrese et al., 2011; Gonzalez et al., 2008; Song et al., 2010). Others initially designed their own monitoring hardware (Garbarino et al., 2014) for getting higher precision and quality of data; these devices were usually personal digital assistants (PDAs) or adapted recorders. Before the disruption of smartphones around the year 2010 worldwide, it is worthy to mention the studies in Mehl et al. (2001) for the analysis of daily voice records, in Bolger et al. (2003) for explaining variability in the diary reports of individuals and in Choudhury et al. (2008) for activity recognition while protecting the privacy of users.

In the case of psychiatric applications, the principal advantage of personal mobile devices is that their embedded monitoring systems are completely unobtrusive. This avoids direct interactions with patients, that are often time-consuming, and limits potential counfounders due to the self-representation. This last point is sometimes the cause of subjective data gathering, strongly dependent of the true sufferer mental state. Moreover, it has also motivated the apparition of electronic mental health (e-Mental Health) protocols, that nowadays is an emergent field (Osmani, 2015; Firth et al., 2016; Barrigón et al., 2017; Larsen et al., 2015; Saha et al., 2016; Marzano et al., 2015). Moreover, over the past years, both traditional ambulatory assessment (AA) and ecological momentary assessment (EMA) methods in clinical psychopatology, initially in the form of paper-and-pencil questionnaires, have accepted the presence of digital systems while keeping the idea of *tracking* patients behavioral dynamics.

Quantitatively, understanding how a mental patient behaves and the characteristic changes reflected in their smartphone metadata, location traces or communication logs has appeared as a fundamental contribution in mental health (Madan et al., 2010), that often require the existence of *ground truth* metrics. In syntomatic individuals, the initial systems required interaction with devices, e.g. text their mood state regularly, which appeared to be problematic in certain scenarios. In Canzian and Musolesi (2015), they mix both answers to questionnaires with the periodic information about mobility and the location of patients. Importantly, it is demonstrated that there exists an underlying correlation between the behavioral mobility dynamics of patients and their mood state. Indeed, they first predict *changes* in the behavioral patterns of sufferers with depression from their mobility data. In this section, we follow the spirit of this last work.

Regarding the comprehension (and modelling) of human circadian dynamics and their adjacent digital phenotypes (Madan et al., 2010; Aledavood et al., 2015a) in affective disorders, the novel pervasive services are ideally suited for capturing the behavioral *states* of patients. Mainly, the objective is to capture their routines during their daily life. Existing approaches have already explored the analysis of human behavior from numerous modalities of information, such as activity recognition from wearable sensors (Sano et al., 2015;

Taylor et al., 2017), communication registers (Aledavood et al., 2015b), app usage (Torous et al., 2018b), text and voice recognition (Yamashita et al., 2019) and more similarly to ours, mobility metrics (Canzian and Musolesi, 2015).

Changes in Digital Phenotypes

The detection of relapses and early intervention systems in mental health diseases have been previously investigated in several existing approaches, mostly focused on its application to schizophrenia (Barnett et al., 2018; Torous et al., 2017). More recent works have already installed specialized *software* on mobile devices and have been successfully used to monitor and improve medication adherence in chronic patients with mental disorders (Barrigón et al., 2017), to detect changes (Berrouiguet et al., 2018) or even for *digital phenotyping*, such as in Onnela and Rauch (2016).

The ideas presented in this chapter are related to four different scopes of research and their applications: i) early detection of changes, often understood as relapses by doctors, ii) human behavior modeling, iii) mobility analysis and, more technically, iv) online learning. In terms of relapse prediction and the detection of significant changes, the method in Barnett et al. (2018) addresses this problem using anomaly detection methods, which finds outliers using a statistical error test. The main difference with the approach discussed in this thesis is that we study probabilistic methodologies for change-point detection instead. In our scenario, this family of detection models has the advantage of handling with high-dimensional and heterogeneous data.

On the other hand, human behavior modeling and digital phenotyping via smartphones have been largely studied since the first works related to *circadian* routine analysis from longitudinal data in Pentland and Liu (1999) and, more recently, in Begole et al. (2003); Eagle and Pentland (2009); Aledavood et al. (2015a). Our approach approximates each individual behavior by means of their mobility patterns, similarly to Canzian and Musolesi (2015), where they first used location traces to monitor individuals affected by depressive mood disorders. Additionally, the understanding of mobility' singular mechanisms in human behavior is studied in Lima et al. (2016); Joseph et al. (2011) or Barnett and Onnela (2020). This last one is also related to our recognition method and learns trajectories from latitude-longituded traces even in presence of missing samples.

- A short discussion on the differences between anomaly and change-point detection is included in Ch. 4.

5.1.1 The Evidence-Based Behavior (eB2) Study

The starting point for the development of an unobtrusive detection system is the modeling of the circadian patterns of mental health patients. To do that, we monitored their daily digital registers thanks to a medical study approved by the Ethics Committee of the Psychiatry Department at the Universidad Hospital Fundación Jiménez Díaz in Madrid, Spain. The digital information is gathered by the Evidence Based-Behavior (eB2) monitoring app, which was also used in the study of Berrouiguet et al. (2018).

Concretely, the *app* under the study collects data from a wide range of sensors equipped in patients' smartphones, for instance, actigraphy, GPS location traces, metadata from the Google location API, app usage logs, registers of nearby WIFI stations and Bluetooh devices. In some cases, if activity recognition is needed, inertial measurements are also gathered. Moreover, the *app* was developed to run in the background and patients do not interact with the system, only during the initial configuration. It was also designed with battery-safe considerations, like non-continuous storage of raw traces, automatic sleep and wake modes, and self-relaunch if the program is shut down by the user or the smartphone is turned off on purpose or accidentally.

- Reference numbers PIC-66/2017-FJD and EO-76/2013-FJD-HIE-HRJC.
- Available at https://eb2.tech.

In this preliminary study, we were specially interested in analysing mobility metrics from location information, which is demostrated to be a primary indicator of the behavioral *state* of an individual, for example, of his/her depressive symptoms (Canzian and Musolesi, 2015), and also allows for easier clinical interpretation. Thus, we used traces of latitude-longitude coordinates irregularly recorded every 3-5 minutes to build reliable representations of the patient's mobility. This degree of temporal precision is sufficient to capture whether a patient is active or inactive during the main stages of a day, that is, morning, afternoon, evening or night.

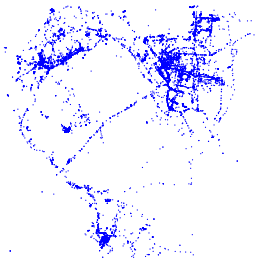


Figure 5.1: Example of the collection of location traces recorded in the initial study for one patient. Longitude-latitude pairs are preprocessed for privacy preservation. Mobility metrics are obtained from the raw data.

In this study, we also proposed to fuse two types of mobility data. One is a distance-based metric, e.g. kilometers travelled from the previous point, and the other is based on the patient's location. This is, whether a patient is near or in some registered place of interest (Eagle and Pentland, 2009). We also assume that the mobility phenotypes for every day can be accurately represented using time-slots of *at least* one hour. These slots could be even reduced to 30-15 minutes if more precision is required in our mobility metrics. However, the reader should notice that higher frequencies yield larger dimensionalities, which in turn, would be in the order of *hundreds* and would also include more missing values.

Raw Mobility Data Preprocessing

Based on Canzian and Musolesi (2015) and Eagle and Pentland (2009), we use the monitored raw traces of latitude-longitude pairs to calculate all distances, in kilometers, between sequential locations. For this task, using the *haversine* formula, we first transform every coordinate into an euclidean projection onto $\mathbb{R}^2$, which is easily measurable. Moreover, to overcome the irregular sampling of input location points, we establish a set of hourly frames $H = [h_0, h_1, \cdots, h_{23}]$, where each $h_j = \{\text{lat}_n, \text{long}_n\}_{n=1}^{N_j}$ collects all latitude-longitude pairs if the timestamp is in that particular frame, e.g. 00:00-00:59 for h_0.

For each partition h_j in a given day t, we compute a real-valued variable x_{jt}, which represents the logarithm of the sum for all *wandered* distances in the particular jth frame. Together, these variables shape the observation vectors $\boldsymbol{x}_{\text{real}} \in \mathbb{R}^D$, that correspond to the total log-distance covered by a patient during every single day. In this case, we assume $D = 24$. Additionally, for the location-based variables, we follow a similar strategy, but instead, we compute distances between every latitude-longitude pair in h_j given a *fixed* location. In this case, the location used is the patient's home or position of reference. If one of the new values is less than a small quantity ϵ, say 50 meters, we may assume that the patient has been in that particular location during that moment of the day. Therefore, we can also define the binary vectors $\boldsymbol{x}_{\text{bin}} \in \{0, 1\}$, where 1 and 0 mean respectively, the *presence* and *absence* at home. Together, both $\boldsymbol{x}_{\text{real}}$ and $\boldsymbol{x}_{\text{bin}}$ are part of the heterogeneous daily representation $\boldsymbol{x}_t$ of a patient's mobility. An example of this multivariate representation is depicted in Fig. 5.5.

5.1.2 Detailed Analysis of Selected Patients

In the initial results presented in Berrouiguet et al. (2018), an unsupervised detection method was performed within a qualitative analysis by clinicians of a first trial sample of 5 patients out of the 38 patients enrolled in the eB2 study between April 6th and December 14th of 2017. The duration of the eB2 study was even larger for 2-years and controlled by the Fundación Jímenez Díaz and particularly, by the clinicians (doctors, clinical psychologists and nurses) from the team led by Dr. E. Baca-García.

Short Description of the Study

The participants recruited in the initial study are sufferers from the outpatient mental health center of the Psychiatry Department in the Fundación Jímenez Díaz, in Madrid, Spain. This department is also part of the National Health Service and provides service to around 420,000 people in the Madrid area. The inclusion criteria for the patients in the study was to be either male or female, more than 18 years old, already diagnosed with mood disorders and coping with depression. Additionally, patients must be own a smartphone with Internet conection and ANDROID or iOS operating system.

Participants were excluded if they were underage, illiterate or enrolled in other studies or clinical trials in the hospital. The participants had to attend, at least, to 2 appointments with doctors before the study. The initial recruitment of the study in 2017 was of 38 patients, and 5 of them were specifically analised with the present behavior change detection tool in Berrouiguet et al. (2018).

The baseline characteristics of the clinical counterpart of the study were, first, to record in-person interviews with clinicians *a priori*. The variables collected during the interviews were sex, age, Patient Health Questionnaire-9 (PHQ-9) score (Kroenke et al., 2001), diagnosis and treatment. The clinical diagnoses were made by psychiatrists and coded according to the ICD-10 scale (World Health Organization et al., 1992) for mental disorders. Thus, during each appointment, the psychiatrist in charge administered the PHQ-9 questionnare to assess depression. These metrics were later introduced into a secured electronic health record. Each patient was identified by a numeric code which ensured her/his anonymity. There were not a control group in this study.

Unsupervised Modeling of Behavioral Profiles

The unsupervised modelling of profiles and the posterior detection of changes from them was comprised in *two* main algorithms. The first one, was the heterogeneous latent class model presented in Ch. 2, Sec. 2.1.2. We applied it according to the mobility metrics described above, and the latent class estimate is what we later assume to be the hidden *behavioral profile*. In the case of this study, the mobility metrics could show, for instance, whether a patient was more active in the morning, afternoon or evening, or even non active at all. The average mobility function during the 24h of a day is clearly seen in Fig. 5.2. Vertical axes are on logarithmic scale, so large mobility values (more than 10km per hour) are likely vehicular transportations, e.g. car driving, train or flights in the case of holidays. However, this last case was not observed in our data collection.

All aggregated distances were stacked in 24-dimensional vectors, and each vector corresponds to one single day per patient. We remark that the observations from different patients were not mixed, so the modelling tool was completely personalized. The inference process of the latent behavior as well as the learning of mean and covariance parameters was performed using the EM algorithm. The mean vector parameters are the ones plotted in both Fig. 5.2 and Fig. 5.3. Results from patients C, D, E were omitted of this thesis for a reason of clarity

in the manuscript. They can be found in Berrouiguet et al. (2018). Additionally, the mean vectors, called *mobility profiles* by the clinicians, were used for the analysis of changes and symptoms of patients.

In a few cases, the preprocessed data contained missing values that we also estimated. For this task, the EM algorithm is also valid due to it can be extended under the assumption that lost dimensions are also latent variables. We used the approach of Ghahramani and Jordan (1994) in all the missing scenarios of the present chapter, for both Bernoulli and Gaussian likelihood densities. Additionaly, one final detail must be signaled. The selection of the order in the latent discrete densities modifies the output prediction of the behavioral states. That is, how many profiles a patient may take during the period under the study. This selection depends on the number of observations taken, e.g. number of days monitored. Hence, we used an automatic selection procedure, based on the minimum description length (MDL) criterion (Murphy, 2012). Later in the next versions of the tool, other selection methods were included, for instance, Bayesian information criterion (BIC) and Akaike information criterion (AIC). In all cases, the maximum number of behavioral profiles rarely was larger than $K = 10$, even having patients with 2-3 years data at the end of this thesis.

Unstable Behavioral Dynamics

Regardless of whether a mental health patient in the eB2 study was *stable* or not, the modelled profiles might suffer changes and variabilities of different type. For example, one *class* may be associated with a profile of higher variance parameter. We named this as *intrinsic* profile variabilities. Moreover, we see that the largest observed behavior also includes changes from day to day owing to weekends or public holidays. This observation led us to assume the main behavioral pattern is not a single profile itself but an arbitrary combination of likely profiles. This idea is the one that makes us different from other similar approaches in the literature, for instance, in Barnett et al. (2018). This understandable as we are agnostic with respect to the evaluation of one profile or another, and our interest lie exclusively on the transition between profiles, not the profile.

Hence, to detect changes in mobility patterns, it is not suffice to identify changes in the sequence of latent behavioral profiles or just unexpected outliers. Instead, we look for abrupt transitions in the generative distribution of these profiles, that is, changes in the natural parameters of the densities. To detect the sudden transitions or shifts, we apply change-point detection methods. Here is where we establish the connection with the *hierarchical* CPD models presented in Ch. 5.

The key properties of the CPD detection framework developed is that it handles i) high-dimensional data, ii) latent class assignments and iii) missing variables in the recorded collections. In this preliminary study, the hierarchical detector is linked to the latent model given mobility data. This approach will be later extended to other types of data registers and presented in the following sections of this chapter.

• As a kind reminder, the word *hierarchical* refers to the ability of the CPD detector to identify changes directly in the generative distribution of the latent variables rather than the high-dimensional data collection.

Initial Assessment and Clinical Description

The pilot study showed us that the methodology could aid clinicians in the mental health practice to detect critical relapses or other clinical changes in an unobtrusive, passive and quick manner. However, one should notice that the events pointed by the detector lack of interpretability, what must be done by an expert clinician. Having said this, we remark that the initial study illustrate the potential applications of the eB2 system in the future, as an assistance tool for patients with depressive disorders.

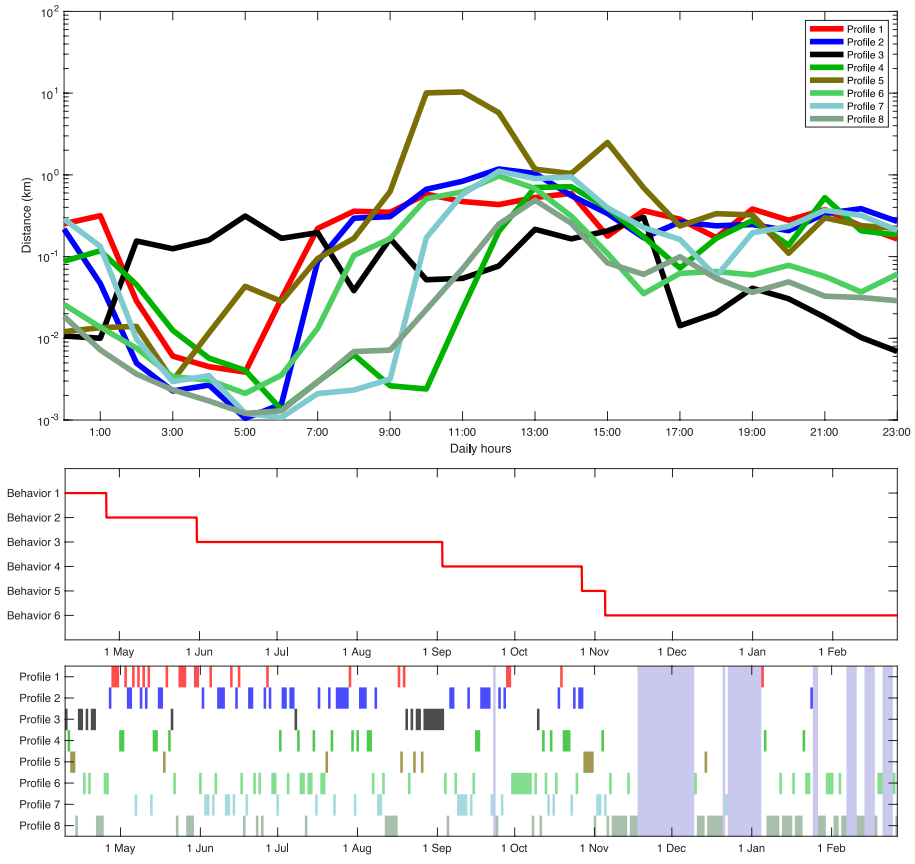


Figure 5.2: (UPPER PLOT) Distance traveled profiles of patient A. The maximum number of behavioral profiles is set to $K = 8$. All profiles except from the *black* colored show highest activity during the diurnal hours, with differences in the intensities and the last and first hour of mobility between the resting period. (LOWER PLOTS) Representation of changes in mobility patterns identified by the detection method and the corresponding features dring the study participation. *Grey* colored bars indicate days that are missing.

In the following lines, we shed light on the findings and results, with a few detailed analyses for patients (A) and (B) and their behavioral dynamics. We remind that the clinical assessment was assisted by the PHQ-9 questionnaires.

• Red horizontal lines in Fig. 5.2 indicate the partitions estimated by CPD.

PATIENT A) – Patient A was a diagnosed 56-years-old woman with recurrent depressive mood disorder and fibromyalgia. She was prescribed with a daily oral medication. She also described regular bedtime and wake-up hours during the medical study. During the previous clinical assessments, the patient showed high scores of depression in the PHQ-Q questionnaire, concretely on dates April 6, 2017 and May 31, 2017. She began her participation in the study on April 6, 2017 and continued until February 28, 2018. Her smartphone device was a Samsung Galaxy S7 which ran on ANDROID system.

In Fig. 5.2, we plotted the mobility profiles as well as the inferred behavioral sequence within a change-point analysis. In the lower plots, we see that *five* events are detected. The

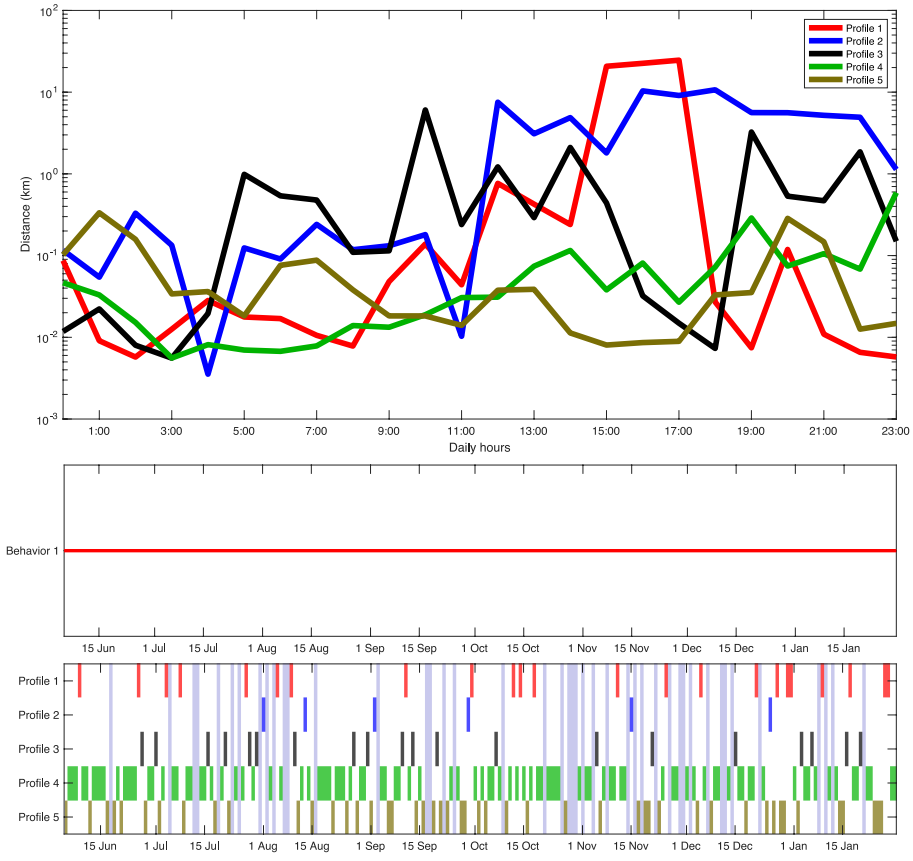


Figure 5.3: (UPPER PLOT) Distance traveled profiles of patient B. Mobility values are approximately below the threshold of 10km per hour. (LOWER PLOTS) Representation of changes in mobility patterns identified by the detection method and the corresponding features dring the study participation. Changes are not identified by the detector, so a single behavioral pattern is assumed to persist along the monitored time.

days between consecutive CPs are $\{17, 15, 95, 54, 9, 113\}$. The days with noted changes were April 26, May 31, August 19, September 3, October 27 and November, 5.

PATIENT B) – Patient B was a 45-years-old woman with diagnosed dysthymia. She was also prescribed with a daily oral medication. The clinical assessment via questionnaires indicated a trend of improvement in her depressive symptoms. Particularly, on June 7, 2017, her PHQ-9 score was 20 and on July 5, 2017 she showed a PHQ-9 of 8. Additionally, her medical records showed an improvement during the evaluations. She initiated her participation in the study on June 7, 2017 and finished on January, 30 of the same year. Similarly as Patient B, she owned a Samsung Galaxy smartphone with ANDROID system.

In this case, the behavior modelling tood identified *five* latent profiles of mobility during the monitored period. Fig. 5.3 shows that the method did not identify any change-point and a single behavioral pattern persisted. The most common distance function indicates that the most likely profile is of low-mobility, (any single hour with a distance travelled >1km). The

findings on this particular patient are correlated with the reduction in the main symptoms expressed in cognitive areas during evaluation.

5.1.3 Principal Findings

The pilot study on human behavior learning from mobility data showed evidence signs that the (initial) proposed method could aid to detect relapses or clinically relevant events. This type of events are indicated by the ML model as change-points. However, a fundamental step for the potential use of the tool in clinical settings of mental health is the task of interpreting the results. We remark that the model facilitates the clinical interpretation by doctors, as the final output consists of latent discrete indicators (*profiles*), associated with higher or lower mobility. The idea is to maintain this sort of structure for an easier comprehension in a medical environment, while at the same time, incorporating new sources of data. We will see this in the following sections of the present chapter.

Clinical Contextualization

Interestingly, the detected events of change and their adjacent profiles represented different clinical scenarios in the study. For instance, patient B showed no change-points, whereas for patient A, the shifts of behavior represent a worsening. In particular, the algorithm detected this worsening on April 26, 2017 (first CP in Fig. 5.2). The PHQ-9 depression score increased between the beginning of the study on April 6, 2017 and May 31, 2017. The participant did not show up for the clinical interview in September 2017. Although, she continued with the protocol and the eB2 monitoring activated, so we cannot establish clinical correlations during that period of the study. A CP event was detected on September 1, 2017, which might be related to the absence, but clinicians did not establish any conclusion.

Additionally, the other three patients in the study (C, D, E) represent conditions of both improvement and worsening, depending on the specific participant. In patient D, the absence of CPs indicates a stability in the symptoms. However, for patient E, a progressive lost of daytime activity profiles was correlated with a clinical worsening. On the other hand, the results obtained from patient B are an example in which the sequence of behavioral profiles did not show any abrupt change, but there was indeed a clinical improvement. Our hypothesis in this last case is that extra sources of information could help in the identification of this improvement, but this analysis was out of the scope of the work at this point of the initial study.

5.1.4 Short Conclusions of the Study

One important conclusion of the system is that both the appearance or disappearance of high mobility profiles could be indicators of the worsening of a patient. This is important for the study, as long as we are interested in the technical contribution of the CPD models, which only indicate the presence of shifts in the underlying densities. Hence, we cannot directly analyze the properties of profiles, something that should correspond to the clinicians. Hence, the purpose of this preliminar work was first, to make an initial implementation of the ML tool within mobility data from patients and second, to obtain an evaluation from doctors about the correlation between CPs and clinical events. Both objectives were successfully achieved. The eB2 study shows that the application of probabilistic modelling within the CPD methods developed in Ch. 4 is feasible, advancing in the implementation of unsupervised ML approaches in mental health.

So far, only location-based data were used, which led to an easier preprocessing stage of the observations, feature design and allowed for a quicker clinical interpretation. However, it is crucial to fuse more behavioral data from additional sources, e.g. activity, social networks, communications, in order to obtain a larger representation of the daily state of chronic patients. As we will show in this chapter, there are specific areas of psychiatric science were this tool has a potential impact for the well-being of sufferers, for example, in suicide prevention.

5.2 Suicide Attempt Prevention

Severe affective disorders with high prevalence, such as depression or bipolar diseases, are mental health illnesses that affect about 2% of the world's population (James et al., 2018; World Health Organization, 2019). In the worst cases, more than one million people worldwide commit suicide every year. Particularly, 33,000 suicide deaths ocurred every year in the U.S. between 2001 and 2009, and it is considered among the top *five* causes of death for adults under 45-years-old in the same country (Office of the Surgeon General (US) et al., 2012). Despite the efforts of national healthcare systems for reducing the number of suicide attempts, it is still a growing problem than requires *effective* and *efficient* methods for prevention. The principal strategies are usually focused on both the detection and treatment of depression by clinicians, something than unfortunately remained unchanged during the first decade of the century (Kessler et al., 2005).

In practice, the degree of risk and disability has been traditionally assessed by clinicians using periodic patient reports, structured questionnaires, e.g. PHQ-9 for assessing depression symptoms, the assistance of caregivers or time-consuming evaluations during periodic appointments. However, risk assessments of suicide involve to consider both *static* and *non-static* factors. In the case of the dynamic factors, we may consider alcohol/drug substance abuse, evolution of mental health disorders, post-hospitalization transition to daily routine, social support, behavioral changes and access to clinical environments (Torous et al., 2018a). Assessing those risks at a single point in time does not capture the dynamics of the patients. Smartphones and new mobile electronic devices, e.g. wearables, offer a new horizon in suicide prevention methods.

In particular, the ubiquitous conditions of smarpthones in the pockets of patients have motivated plenty of advances in the *passive* assessment of mental health sufferers with high suicide risk factors. Examples are Torous et al. (2015); Berrouiguet et al. (2019) and Robinson et al. (2016) in the analysis of social media. In this context, we remark the apparition of new techniques in psychiatric science based on the idea of evaluating patients out the hospitalary areas within computerized methods. This family of methods are often referred to as *ecological momentary assessment* (EMA) (Stone et al., 2007), and they aim to capture symptoms at the moment they occur very shortly thereafter. This spirit is also seeked in the recently new technologies developed for suicide attempt prevention.

5.2.1 Machine Learning for Suicide Events

However, it does not suffice to only monitor and record real-time variables from the daily lifes of patients. The second step for accurate suicide prediction and prevention is the capability of these new digital systems to analyze large collections of data. The final objective is to generate easily interpretable summaries of the risk factors that patients are suffering via machine learning and stastistical modelling. Several methods already achieved this goal in a supervised manner. For instance, in Simon et al. (2018) authors use more than 300

demographic variables recorded from the previous 5 years prior to the first suicide attempt as the input data to a penalized logistic regression model. Another example, presented in Barak-Corren et al. (2017), uses a similar setup to predict suicidal behavior directly from electronic health records (EHR) with years in advance. Finally, and more similar to the sequential scenarios considered in this thesis, Peis et al. (2019) introduces recurrent neural networks (RNNs) to predict suicidal ideation from both EHR and EMA data.

In this section, we extend the detection model presented in Sec. 5.1.1 to develop a novel ML methodology for the automatic assessment of daily behavioral features of mental health patients with at least one suicide attempt. Also, we present an early detection tool of the behavioral instabilities captured by the latent variable model. The smartphone-based system is both *passive* and *personalized*, similarly as in the eB2 study previously described. It consists of three main blocks depicted in the diagram of Fig. 5.4. These ones are:

i) the analysis and pre-processing of *app* usage and mobility raw data from patients' mobile devices,

ii) the Bayesian modelling of high-dimensional observations within the heterogeneous likelihood functions of Ch. 2, for obtaining discrete latent indicators of the daily behavioral profiles of patients (as in Sec. 5.1.1),

iii) the *online* detection of change-points in the low-dimensional sequence of latent behavior identifiers.

The output of the system is validated within the clinical data provided by clinicians from urgencies and hospital interventions. The validation results are presented in the last section of this chapter.

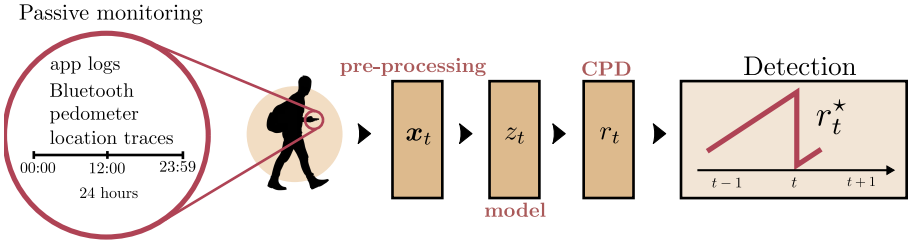


Figure 5.4: Illustrative diagram of the proposed mobile-based tool for the early detection of behavioral changes in the context of suicide attempt prevention. CPD is the acronym of change-point detector. The $\boldsymbol{x}_t$ shaded box represents the process of stacking data attributes for building up the high-dimensional observations from raw data. The *saw-tooth* red line indicates a change at time t.

5.2.2 Mobile Health Data

We have already reviewed in this chapter how mobile electronic devices, particularly smartphones and wearables, have a pervasive presence in our daily lifes. Hence, we can define the behavioral *footprint* as the implicit projection made by us into the digital data that we generate at every moment. In this secondary stage of the behavior learning approach for

mental health, we are mainly interested in this behavioral footprint. To extend the preliminary study in Berrouiguet et al. (2018), we now take advantage of this ubiquity, and all the available resources recorded by the eB2 monitoring *app* of Sec. 5.1.1. The idea is to build now high-dimensional representations of one day, stacking more longitudinal data that the ones captured only via mobility metrics. We remind that the medical study allowed (under the supervision of clinicians) periodic records of actigraphy sensors, anonymized location traces, *app* usage logs, nearby WiFi stations and seen Bluetooth devices.

Notice that stacking new data features from several sensors turns the modelling problem into the heterogeneous scenario, presented in Ch. 2, as long as we now have a multivariate problem with different types of statistical variables. However, we maintain our strategy of learning a latent discrete structure, whose interpretability for psychiatry clinicians is easier and fits particularly well within the surrogate CPD method.

Longitudinal daily representations – The first stage on the process of design the heterogeneous latent variable model begins now with the pre-processing of the daily behavioral registers from the monitored patients. The sensed data has been stored using time-slots of half-hour in this case. This precision for the data collection is chosen for having $\sim$200 features per observation vector, which is of higher dimensionality than the model in the previous medical study. To obtain a reliable representation of *one* day or time step t in the sequential scenario, we choose two behavioral areas: mobility and social activity. For the mobility features, we use the *steps count* and the *location traces*. The last subset of variables are the ones considered in Berrouiguet et al. (2018). We also compute the *distance travelled* between consecutive points as we did in the previous Sec. 5.1.1. Then, we generate two real-valued vectors, one subset per data-type, $\boldsymbol{x}_{\text{real}} \in \mathbb{R}^D$, that represents the total log-distance and log-steps monitored. In this case, we considered $D = 48$, so we increased the number of slots per day.

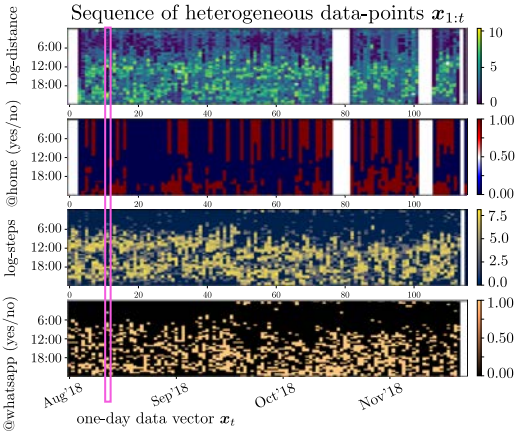


Figure 5.5: Multi-sensor daily representations of the heterogeneous data gathered by smartphones.

For the social activity variables, such as *phone usage* and *home-presence indicators*, we obtained two binary vectors, $\boldsymbol{x}_{\text{bin}} \in [0, 1]^D$, where 1 indicated the mobile-phone usage and presence at gome, respectively. The metrics related to *app* usage are based on the ones considered in the experiments for heterogeneous GP models in Ch. 2. Together, the continuous and discrete variables, $\boldsymbol{x}_{\text{real}}$ and $\boldsymbol{x}_{\text{bin}}$, compose the heterogeneous daily representation $\boldsymbol{x}_t$ of a patient. An example of the sequence of high-dimensional observations is shown in Fig. 5.5, where *white* bars represent the missing features in some of the features. This partial observation problem will be trated in the following section. Finally, we must remark the *circadian* pattern of the patient, visible accross the different statistical variables. We also referred to it as the behavioral footprint in the smartphone generated data.

5.2.3 Passive Detection of Behavioral Shifts

In this scenario, to early detect abrupt changes in the sequence of behavioral observations from psychiatric patients, we also project the observations in a lower-dimensional manifold $\boldsymbol{z}$. Thus, we are interested in converting the high-dimensional sequences $\boldsymbol{x}_{1:t}$ into interpretable discrete representations using the latent class model in Sec. 2.1.2. For this task, we infer the underlying sequence of discrete indicators $\boldsymbol{z}_{1:t}$ from the heterogeneous variables presented above. Notice that this generative model corresponds to a Gaussian-Bernoulli mixture likelihood with ∼100 dimensions per statistical data type.

Heterogeneous Behavior Modelling

In this stage of the tool, we infer the latent categorical variables $\boldsymbol{z}_{1:t}$ that represents the *type of routine* or *daily circadian profile* of the patient given the heterogeneous data collection. Based on the type of *circadian models* described in Moreno-Muñoz et al. (2020) and in Sec. 2.1.3, the individual likelihood terms $p(\boldsymbol{x}_{\text{real}}|\boldsymbol{\theta})$ and $p(\boldsymbol{x}_{\text{bin}}|\boldsymbol{\theta})$ are multivariate Gaussian and Bernoulli densities, respectively. We now build the covariance matrices of the Gaussian distribution from the periodic non-stationary *kernel* functions in Sec. 2.1.4. The idea is to express the short-term correlations, e.g. between daily hours, with this sort of periodic mappings. If we assume that there are a maximum number of K behavioral profiles $z_t \in [1, 2, \cdots, K]$, generated *a priori* as $z_t \sim \text{Cat}(\boldsymbol{\pi}_t)$, then the vector observations are given by

$$\boldsymbol{x}_{\text{real},t}|z_t \sim \mathcal{N}(\boldsymbol{f}_k, \boldsymbol{\Sigma}_k), \tag{5.1}$$

$$\boldsymbol{x}_{\text{bin},t}|z_t \sim \text{Ber}(\boldsymbol{\mu}_k), \tag{5.2}$$

where $\boldsymbol{\mu}_k \in [0,1]^{2D}$, with $D = 48$, and $\boldsymbol{\Sigma}_k$ as defined in Sec. 2.1.4. Thus, the likelihood distribution model considered takes the form

$$p(\boldsymbol{x}_t|z_t, \{\boldsymbol{\theta}_k^1, \ldots, \boldsymbol{\theta}_k^M\}_{k=1}^K) = \prod_{k=1}^{K}\prod_{j=1}^{M} p(\boldsymbol{x}_t^j|\boldsymbol{\theta}_k^j)^{\{1|z_t=k\}}, \tag{5.3}$$

where M is now the number of heterogeneous components. To infer the sequence of latent class variables as well as model parameters $\{\boldsymbol{\theta}_k\}_{k=1}^K$ and the prior density hyperparameters, we use the EM algorithm. Furthermore, the model also handles partially missing observations and even unobserved features. We point out the three existing types of missing information in the behavioral data from patients:

1) MISSING HOURS – Given the multivariate observation vector $\boldsymbol{x}_t$, it is already possible to find missing components randomly, e.g. time slots with no location data due to the signal lost or low battery. To simplify the exposition, we split each sample as $\boldsymbol{x}_t = \{\boldsymbol{x}_t^o, \boldsymbol{x}_t^m\}$, where missing ($m$) values have a dimensionality $\dim(\boldsymbol{x}_t^m) \leq J$ and for the observed (o) ones, the dimensionality is $\dim(\boldsymbol{x}_t^o) = J - \dim(\boldsymbol{x}_t^m)$. Here, we use J as the total dimension of the daily vector $\boldsymbol{x}_t$. Under the assumption that the model is MCAR (Rubin, 1976), we follow the approach of Ghahramani and Jordan (1994) in this scenario. Hence, the missing values $\boldsymbol{x}_t^m$ are treated as latent variables at each time-step t and iteration of the EM algorithm.

2) MISSING REGISTERS – A practical example of this sort of missing features can be seen in Fig. 5.5, where the binary home-presence indicators and the log-distance metrics are lost at a few consecutive time-steps. This is understandable, as we fuse longitudinal data from multiple sensors and information sources that are typically independent. In some cases, if the signal is lost, location traces become *null*, but others as the pedometer and some *apps* in

the cellular might continue working. In this case, we do consider the same approach as in the previous example 1). Hence, a latent profile indicator z_t is still obtained and lost registers are treated as latent variables that the EM algorithm estimates and interpolates at every iteration.

3) MISSING DAYS – A subtle difference from the previous case is that now an entire day t may be completely unobserved, i.e. the smartphone is turned off, leading to an entire lost observation $\boldsymbol{x}_t$. The problem comes when one tries to apply EM in this case for the interpolation of values, which is not possible as the expected values are considered to be conditioned on the observed features. Thus, if all features in the vector $\boldsymbol{x}_t$ are lost, then, performing expectation operations does not work. Then, instead of addressing the estimation of unobserved values $\boldsymbol{x}_t^m$ in the heterogeneous mixture model, it is preferable to directly consider the latent class assignment z_t as *missing*. In this case, we apply the CPD variant in Sec. 4.2.3. An example of these lost days is shown in Fig. 5.6 as *white* bars in the upper plot.

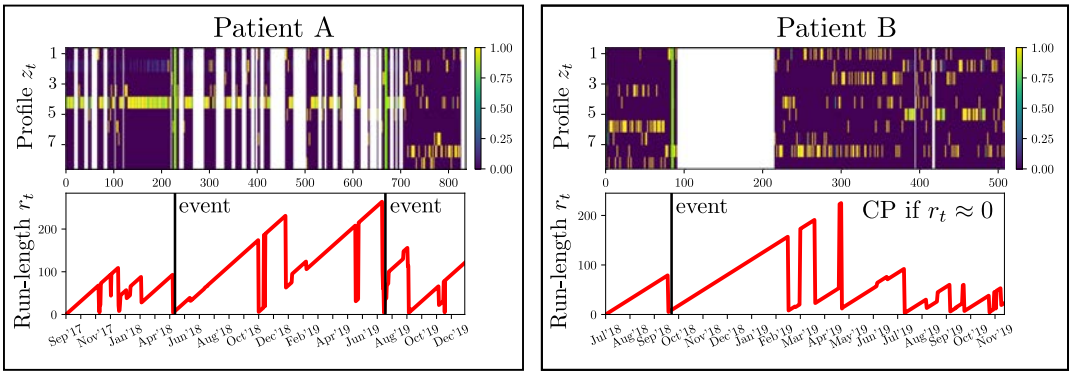


Figure 5.6: Analysis of behavioral changes for two patients in the studied population. Vertical green and black lines indicate the *true* clinical events. Upper plots show the probabilities of each behavioral profile per day. White spaces indicate missing days. Patient A has a likely routine profile ($k = 4$). Patient B turned off the smartphone after the clinical event.

The final goal of this stage is to obtain the posterior probabilities $p(\boldsymbol{z}_{1:t}|\boldsymbol{x}_{1:t}, \boldsymbol{\theta})$ given the heterogeneous observations, that we will later use for the detection of the behavioral change-points in the next section.

Detection of Behavioral Shifts

We take the behavioral output sequence of circadian profiles $\boldsymbol{z}_{1:t}$ for a certain patient, and we aim to directly estimate the abrupt transitions in their generative distribution. An example based on the previous definitions would be to think about a patient whose most likely routines combines "high-mobility" with periodic social interactions. Suddenly, if this profile becomes less usual and it is alternated with several days of "low-activity" across all registers, we may consider that proportions are inverted. In the context of mental health diseases, we must be aware of these transitions as they might be warning signs of an imminent crisis. These are the change-points that we aim to detect, as we did in Sec. 5.1.1 with the mobility data.

In this case, the hierarchical CPD model adopted is the extended version in Romero-Medrano et al. (2020), that we developed in the previous Ch. 4 (in Sec. 4.2.5). The main assumption is that we cannot observe the true discrete sequence $\boldsymbol{z}_{1:t}$ of circadian profiles z_t, but its posterior density $p(z_t|\boldsymbol{x}_t)$ previously computed. Then, based on Romero-Medrano et al. (2020), we draw S i.i.d. realizations of the latent variable, such that

$$z_t^{(1)}, z_t^{(2)}, \ldots, z_t^{(S)} \sim p(z_t|\boldsymbol{x}_t, \boldsymbol{\theta}), \forall t,$$

to fully characterize the probability over the discrete latent classes. This process leads us to consider a multinomial probability model for the CPD, where we define the associated counting vectors as $c_t \in \mathbb{Z}_+^K$. Each t-th component comes from $c_t^k := \sum_{s=1}^{S} \mathbb{I}\{z_t^{(s)} = k\}$. Further details on the inference process can be found in Sec. 4.2.5. Moreover, this multinomial model has shown to increase the precision rate and to reduce the delay in the detection while keeping low computational cost, which is a key property for the considered application.

We continuously infer the posterior probability $p(r_t|\boldsymbol{z}_{1:t})$ at each time step t, where r_t is the *run-length* variable in the CPD method. We remind that it counts the number of steps since the last CP occured. As a consequence, we obtain a measure of uncertainty of the last CP location given the sequence of profiles until that particular moment. For example, the posterior density $p(r_{150} = 5|\boldsymbol{z}_{1:150})$ would indicate the probability of observing a change in the underlying distribution of profiles happened 5 days ago, so at $t = 145$.

Having the posterior densities $p(r_t|\boldsymbol{z}_{1:t})$ for every patient and time-step t, we define a mechanism for detecting shifts between different behavioral dynamics. We use the sequence of *maximum-a-posteriori* (MAP) estimates $r_t^* = \arg\max p(r_t|\boldsymbol{z}_{1:t})$, that are shown in red in the Fig. 5.6. They represent, for a particular t, the most likely day in which the behavior changes. From its proper definition, the optimal values r_t^* takes values from 0 to t. Importantly, we only consider that a behavioral change is detected at time step $t = t'$ if there is an abrupt decrease from $r_{t'-1}^*$ to $r_{t'}^*$. Based on the experiments, we set $r_{t'}^* \approx 0$ as the condition for the detection.

In Fig. 5.6, we included two demonstrative examples of the behavior change-point detection. In the upper plots, we can see the probabilities of belonging to one behavioral profile or another per day. In the left-hand case, there is a likely profile $k = 4$ that dominates during the first two years (until $t = 700$). In the meantime, there are also apparitions of unseen profiles that the CPD indicates. It is also interesting to analyze how the combination of behavioral profiles changes after the second event. In the case of Patient B, there is an event next to $t = 100$. The reader should notice the presence of white bars in the latent profile estimation. This is due to the smartphone was turned off after a suicide attempt, probably due to an hospitalization. Thus, the lack of monitored data avoid any prediction since attributes are not available. Importantly, the principal event was detected *a priori* within a week. The *noisy* behavior of the red CPD curves depends on the setup of the hyper-parameters. We provide insights of the better configuration of the tool in the next section within the experimental results on validation.

5.3 Clinical Validation of Events

In this section, our purpose is to validate the accuracy of the smartphone-based ML tool for detecting the behavioral changes in mental health patients with *at least* one suicide attempt. The experiments carried out typically consists of three parts: i) the modelling of behavioral profiles from the heterogeneous data with missings, and the analysis of the sequence of posterior probabilities $p(\boldsymbol{z}_{1:t}|\boldsymbol{x}_{1:t}, \boldsymbol{\theta})$ as in the upper plot in Fig. 5.6, ii) the application of the CPD method, which in this case is the Multinomial extended version from Sec. 4.2.5, and iii) the characterization of false-alarm and sensitivity rates from the timestamps of events provided by clinicians. The data collected from patients for this clinical validation of the CPs comes from a medical study that is described in the next section.

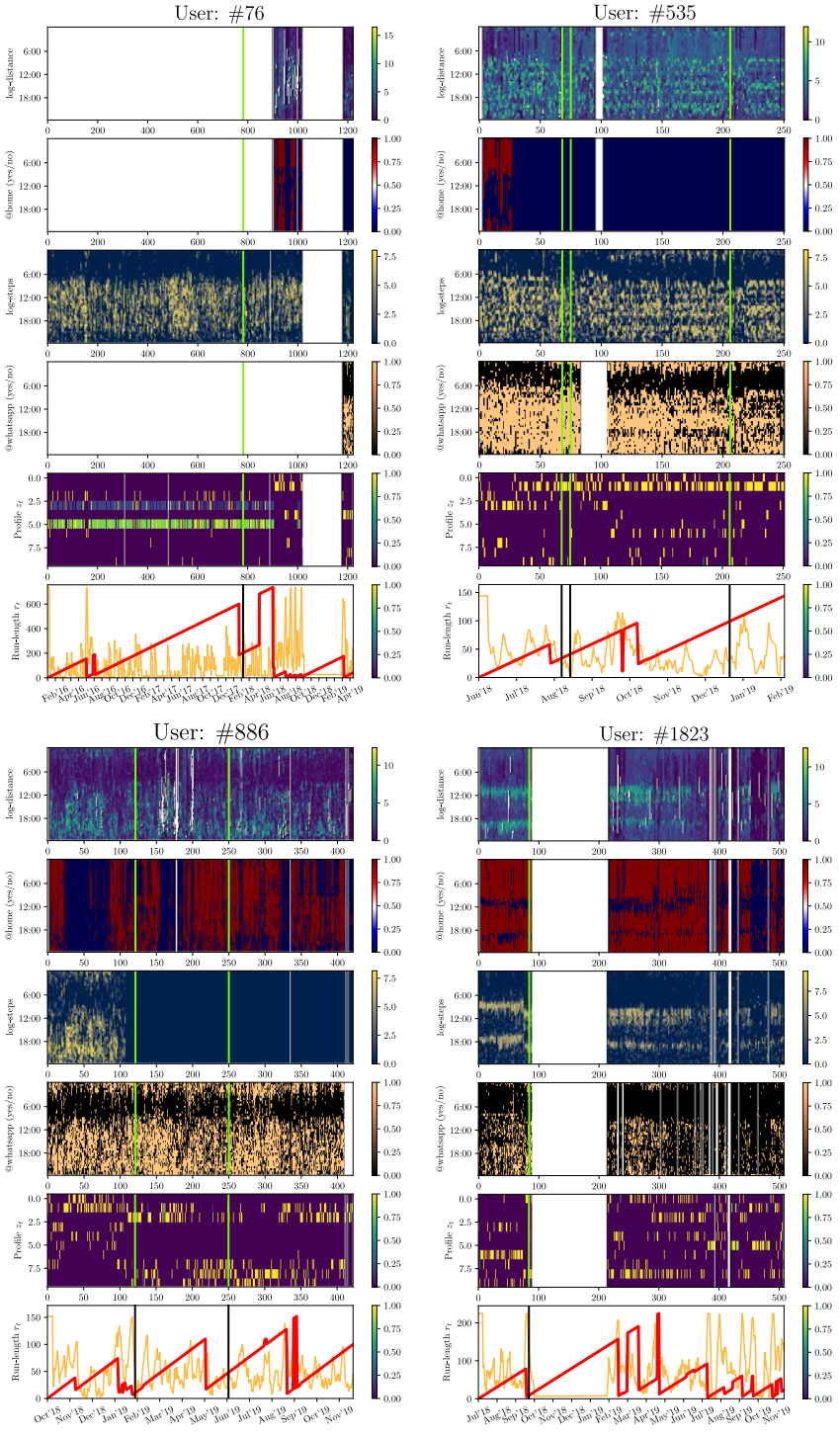


Figure 5.7: Scheme of the data detection pipeline for *four* patients in the collection. From left to right, up to down: {Patient #76, Patient #535, Patient #886, Patient #1823}. Green and black vertical lines indicate the *ground truth* of critical events, e.g. suicide attempt or intervention in urgencies. Patient #1823 turned off the smartphone after incoming in the hospital.

5.3.1 The Smarcrises Study Protocol

The *Smartcrisis* study (Berrouiguet et al., 2019) is a cross-national comparative study with outpatients from the Psychiatry Deparment in the Hospital Fundación Jiménez Díaz (Madrid, Spain) and the University Hospital of Nimes (France). The objective of the study is to screen the digital footprints of mental health patients for their clinical management and to determine the relationship between wish to die, suicide attempts as well as sleep quality.

The eligibility of the patients for the study was similar to Berrouiguet et al. (2018) and builds on the results provided by previous psychiatry studies in the application of e-Health tools for suicide assessment (Husky et al., 2014; Berrouiguet et al., 2014). In this case, patients were assigned to intervention or control groups if they had a personal history of suicide attempts or not. Patients with a current suicidal behavior disorder (according to DSM-5), are always assigned to the intervention group. These ones have had a suicide attempt during the last year. If there is no clinical history of suicide, patients in the study were assigned to the control group.

• The DSM-5 is the "Diagnostic and statistical manual of mental disorders", fifth edition, traditionally used in psychiatry medical practice.

The patients in both groups were monitored by two smartphone *apps*. One collects EMA data (Barrigón et al., 2017) and the other is the same one used in the EB2 study. The initial proposal for recruitment included 1000 patients between the two cross-national sites. The main idea was to capture around 100 suicide re-attempts during the follow up period with the monitoring systems.

Dataset Description

The final data used for the validation consists of a total of 301 outpatients clinically diagnosed during an average period of 346 days per patient. The input attributes to the CPD model are the ones described in Sec. 5.2.3 previously. The total missing rate was 29%, with a rate of 25% missing values per patient. The total number of days monitored in the collection was ∼104k, and the largest register had 1492 days, a bit more than 4 years. These cases longer than the *Smartcrisis* study (which began in 2018) are due to the EB2 app makes queries to third-party installed software, e.g. of fitness type, GOOGLE maps, or the system memory. Sometimes, these ones return data from additional moments out of the follow-up period, which is also pre-processed for privacy preservation. Importantly, the total number of events attempts ocurred during the monitored time was 111. The dates correspond to two types of suidical events: i) registered attempts and ii) urgency interventions due to crisis or self-harm. These dates are the ones used for the validation of CPs. Examples can be found in Fig. 5.6 and Fig. 5.7, where are signaled in green and black-colored curves.

5.3.2 Performance Characterization Metrics

The detector achieves an *area-under-the*-ROC (AUROC) metric of 0.71 in a completely passive and unsupervised manner. The CPD error metric of the tool is significantly over the random guessing threshold. A key aspect of the approach is that a minimum of false-alarm rate is required, due to the psychological costs to sufferers and caregivers in this clinical scenario. This point causes that higher rates of detection are difficult to achieve. Additionally, we remark that some of the suicide attempts in the collection are undetectable, since there are cases where the timestamps correspond to missing dates. We remark that these drawbacks make the obtained AUROC metric even more promising.

The tool can be also personalized to higher or lower probabilities of CP via the λ hyper-parameter of the CPD method. Particularly, this parameter regulates the conditional prior density $p(r_t|r_{t-1})$. We recommend to revisit Sec. 4.2 for this aspect. In addition, we choose

a window of one-week as the warning period, that is, a CP at time-step t placed six days or less before a suicide attempt is considered a *true positive*. This can be also tuned to have longer or shorter detection windows if required. In Fig. 5.8 we show three examples of the characterization curves for the CPD model. Both the number of behavioral profiles K and samples S have a strong aconditioning on the computational cost of the statistical method, so extremely high values could not be considered.

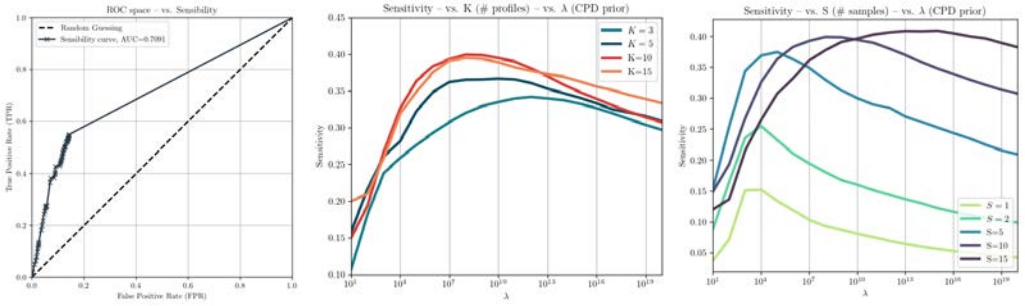


Figure 5.8: Characterization curves of the behavior CPD model. (LEFT) ROC and AUROC value under the curve. We integrated both hyperparameters in the circadian mixture model and the hierarchical CPD method. (MIDDLE) Evolution of the sensitivity rates w.r.t. the number of behavioral profiles considered and the probability of having a CP, λ. (RIGHT) Evolution of the sensitivity w.r.t. the number of samples used in the multinomial version of the hierarchical CPD method and λ.

5.4 Discussion

This chapter developed several version of an unobtrusive ML system that can capture data from the native sensors of patients smartphones. The preliminary results within the EB2 study and its *app* showed the feasibility of unsupervised detection methods in the context of human behavior for mental health. This initial work also confirmed the insights shown by previous studies related to mobility data, that we also aimed to extend. Using patients data and the clinical registers provided by clinicians in the *Smartcrises* study, we introduced a novel extension of the CPD tool for the detection of behavioral shifts with application to suicide attempt prevention. In this case, we built high-precision representations of the mobility and social activity from their personal smartphones. The technical advances presented in the preceding Ch. 4 are put into practice in the detection of CPs from the sequence of discrete behavioral indicators. The use of the multinomial type of CPD model guarantees the robustness of results. Validation experiments on a population of ∼300 psychiatric patients shed light on the potential use of this sort of detection methods in a clinical scenario

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

The preceding chapters developed a detailed analysis of *four* main areas of study. Three of them (Ch. 2, 3 and 4) presented key technical contributions, particularly on advanced statistical methods for different problems in the context of human behavior learning. One last chapter (Ch. 5) put on practice the core advances achieved for the application of probabilistic models in mental health. In this last case, we even considered fundamental contributions to medical problems, as the task of suicide attempt prevention. In this last chapter, we survey the principal contributions of the doctoral thesis, outlining the main results obtaining and a detailed presentation of the main points for future research development.

6.1 Summary of Methods and Contributions

Human behavior learning from smartphone generated data, at this moment the best digital representation of our circadian routines, presents several challenges that we have faced in this thesis. During the doctoral project, we have identified three of these problems that limit direct applications of state-of-the-art statistical methods on them. The first problem is the stochastic component of our behavior and hence, of the data. Here is where probabilistic models make their apparition, as long as they can be robust to wide variations in the high-dimensional observations and, at the same time, maintain a well-fitting of the model for future prediction. In addition to the existing variabilities in the behavioral records, we believe that the assumption of regularly sampled data with *homogeneous* characteristics is not realistic in this context. The blessing of the smartphone's unobtrusive monitoring also leads to some drawbacks. In our case, we are concerned with the flow of observed data from differently engineered sources (and sensors), from mobile devices whose *Internet* connection or power load is not always guaranteed. This idea makes us to consider an *irregular* data scenario, where data might be *heterogeneous*, that is, of different statistical data types, either discrete or continuous. Also we consider a notorious presence of missing features and even missing observations, that cannot be ignored by the model if we want a certain quality in the performance for medical applications. Together, this *irregular* data conditions shape the second problem that we identified in this thesis. Finally, the third challenge is related to the availability of the observed objects. Motivated by mental health applications, we consider that massive storage of sensitive data and long-time delays for producing predictive results is not possible in our context. This consideration opens the door to the adaptation of probabilistic models to both *distributed* and *continual* scenarios. We say distributed data as long as we expect to deploy statistical methods that do not require to access data in a centralized manner, e.g. perform computation directly on the smartphone, for preserving privacy of patients and also faciliting the data preprocessing stages. The term *continual* refers to the particular case of *online* statistical methods that allow for recursively updating models in a sequential manner while at the same time accept the apparition of new unseen tasks. This last topic is also connected with the final technical contribution of this thesis. We are not only interested in modeling human behavior, but also in detecting abrupt transitions, e.g. change-points or significant parameter shifts, under the aforementioned *irregular* data

setups.

We examine these themes, which are specifically fitted to the particular problem of human behavior learning but are also general to the actual settings in modern machine learning. The first part of the thesis considers probabilistic methods, e.g. latent variable models, for the problem of *irregular* data, and particularly, introduces several methods based on heterogeneous likelihood densities. This is deeply explored in the context of Gaussian processes. We formulate the first multi-output GP model that addresses different likelihood models per output function. The model is scalable thanks to the application of approximate inference methods, particularly variational ones. Moreover, the ideas used for GPs are extended to other areas of study, and we also provide a set of auxiliary properties to adapt latent class models to heterogeneous data with 24h periodic components. These works are the central contributions of the Ch. 2.

The second part of the thesis, which is focused exclusively on inference methods, begins by developing a new inference scheme for Gaussian processes where the training data collection is assumed to be distributed. In this case, the model is based on the idea of *recycling* already trained GPs. Applying properties of Gaussian marginals within the infinite-dimensional integral operators, that are needed for building lower bounds in GP models, we are able to build different global models (or ensembles) from the locally trained ones. This approach allows to save models rather than data once these are well-fitted. Then, without the need of revisiting any sample, new GPs can be directly obtained. It is empirically evaluated with state-of-the-art methods and its precise performance announces promising advances on this type of inference. In the Ch. 3, we continue with a similar setup, but now oriented to the sequential learning problem. Particularly, the heterogeneous MOGP approach is extended to accept collections of streaming data. This *continual* GP model is presented for both single-output and multi-output scenarios. We define a new inference mechanism where inducing points, that are typically assumed to be fixed once the sparse GP approximation is fitted, can be re-located and even augmented in a recursive way. This adaptation keeps the encoded uncertainty in the parameters, while at the same time, improves prediction in new areas of the input domain.

Turning to the human behavior applications, the final part of the thesis is focused on the problem of *change*, which is technically developed in Ch. 4 and its applications to mental health reviewed in Ch. 5. This part of the thesis begins with the presentation of the change-point detection problem. Despite that this family of statistical methods is well-known and studied in the literature, particularly in signal processing, we consider Bayesian approaches where new likelihood models can be introduced. Then, we analyze the main drawbacks that sequentially estimating parameters may lead to if we consider the high-dimensional *irregular* data previously introduced. The resulting method introduces *hierarchy* into the change-point detection formulation, developing a consistent latent variable model in the underlying predictive mechanism of the method. Connected with the previous advances on *irregular* data and particularly on heterogeneous observations, the hierarchical CPD algorithm now makes possible to detect change in multivariate sequences of different statistical nature. Several extensions for improvement are also applied to this model, which is finally used as a medical tool in the last chapter of the thesis.

Regarding the application side of the CPD algorithm, we dig into the medical problem of detecting change from behavioral data. After presenting the main strategies for the passive assessment of mental health patients with chronic affective disorders, e.g. depression or bipolar diseases, we provide the details about the preliminary eB2 study. In this medical scenario, we first applied the CPD algorithm over mobility data from raw location traces recorded by smartphones. This initial study sheds light on the possibility of detecting abrupt transitions in the behavioral patterns of patients during their daily life, with a potential impact on

their quality of life. Based on this initial approach, the work is extended for suicide attempt prevention, where the heterogeneous likelihood functions are put into practice within the hierarchical CPD model previously developed. Finally, a validation study of the detected events within the ground truth from interventions in hospitals and urgencies is provided. Importantly, the characterization of the tool demonstrates the viability of the method to be deployed in clinical scenarios.

6.2 Suggestions for Future Research

We conclude this thesis by discussing future directions and open research problems that have not been considered yet in the literature and are also related with the doctoral project. As long as the thesis provides contributions in three main technical themes (heterogeneous data models, GP inference and CPD methods) and one more applied to mental health, we briefly survey the potential implications of models for new domains in the following sections.

6.2.1 Heterogeneous Likelihoods

The heterogeneous MOGP model in Ch. 2 links *at least* one output function to each parameter of the likelihood densities considered. In cases where such likelihood functions requires more-than-one output function, e.g. heteroscedastic Gaussian distribution or the *chained* Poisson likehood model in Saul et al. (2016), a portion of the model is more conditioned by their parameters than others. If we want to consider, for instance, larger categorical densities with $K>1$K parameters, then other linking schemes must be assumed. Despite the problem of modelling categorical data with GPs is currently under study in the literature, we find an interesting application of Ruiz et al. (2018) in this context. More generally, one can introduce the latent variable augmentation to alleviate the computational cost and hence, reduce the number of latent output functions in the likelihood density. If this strategy turns to be successful, then the heterogeneous MOGP could become ubiquitous in machine learning modelling, for example, in classification of documents, language models or recommendation systems. Additionally, in future work, it would be interesting to employ convolutional processes (CP)s as an alternative to the LMC in the multi-output GP prior. Also related with the GP setup, we may consider to automatically discover the heterogeneous data types in advance (Valera and Ghahramani, 2017). It would avoid to type hand-made definitions of the likelihood densities *a priori* and would consists of an input block in the main setting of the tool.

6.2.2 New Perspectives for GPs

In Ch. 3, we have developed two novel inference processes for sparse GP models. In future work for the distributed setting, it would be interesting to extend the GP prior to accept convolutional kernels Van der Wilk et al. (2017). The idea is to introduce convolutional structures for a better characterization of the input-space, as it is done in neural nets. This method also accepts inducing inputs and preserves the variational sparse approximation. In terms of distributed inference and model *recycling*, its contribution to image processing could be much higher. Moreover, the framework could also adopt the functional regularisation in Titsias et al. (2020) for continual learning applications. Precisely, we find that the continual GP in Sec. 3.2 has important connections with other recent approaches in variational inference methods. For example, with Ruiz and Titsias (2019) and their contrastive divergence (VCD) based on *three* KL divergence terms. This idea of a triple regularized bound also emerges

naturally in our continual learning scenario, particularly from the Bayes rule when we avoid revisiting data. Future research lines on this side are, to substitute the multi-output GP prior as in the case of the heterogenous model and even consider non-linear mappings as the mixing operator between latent functions. Moreover, the continual single-output method could be used as the latent baseline in the multivariate time series imputation of Fortuin et al. (2019, 2020). They use a GP to capture temporal dependencies between real-valued latent variables that are later connected to a deep sequential variational autoencoder (VAE) Kingma and Welling (2013). Another promising strategy would be to study the need of increasing the number of inducing inputs M as the input domain augments sequentially. It could be firstly specified via the recent bounds for sparse approximations proposed in Burt et al. (2020). Finally, we may adapt both the single and the multi-output continual model to accept non-stationary latent functions similarly to Zhang et al. (2019) or even an infinite number of them via mixture of experts (Pradier and Perez-Cruz, 2018).

6.2.3 Latent Structures for CPD

Ch. 4 develops new adaptations of the Bayesian CPD algorithm where latent variable models can be introduced. Due to computational purposes and the problem of interpretability in clinical scenarios, we focused primarily in the case of discrete latent variables. However, the hierarchical CPD framework also accepts new latent variable structures, e.g. feature-based as in Griffiths and Ghahramani (2011). The key issue to overcome is the sequential update of the inference, which in the case of latent variable models is not easy. We believe that a first step forward would be to consider binary cases, and then, augment the representation to accept even vector components. Notice that this idea also preserves the interpretability properties that are fundamental for the quick comprehension by clinicians. Interestingly, in Sec. 4.2.4 we made an effort for obtaining an unbounded CPD method in the dimensionality of the latent structure. This would be also required if other models are considered. Finally, we also propose factorial models, highly non-linear parametrizations (e.g. neural nets as in Rezende et al.) or even disentangled variables.

6.2.4 Behavior Modelling in Mental Health

While we have focused on modelling tasks for *irregular* data, the statistical methods developed in this thesis can be also applied to multiple medical scenarios. As we discussed in Ch. 5, there is a significant interest for deploying unobtrusive detectors in the smartphones of patients with severe mental health disorders. In particular, we developed a passive detection tool for suicide attempt prevention, that in our case, recorded data from *four* registers, (log-distance, steps, presence and *app* usage). Importantly, there is a clear evidence that the larger daily representations are, the better the characterisation of behavior is. Having said this, it would be interesting to extend such representations via newer data sources and sensors of several time scales. If privacy is well preserved, adding logs of app usage, communications or text messages could be of key importance. Language models have been also applied in the mental health context, and future evolutions of the tool could also introduce these advances.

APPENDIX A

HIERARCHICAL CHANGE-POINT DETECTION

A.1 Derivation of Run-length Posterior Distributions

The factorization of the joint probability distribution $p(r_t, \boldsymbol{z}_{1:t}, \boldsymbol{x}_{1:t}, \boldsymbol{\theta}_t)$ is recursive, and is based on the original derivation of Adams and MacKay (2007). We first expand the joint distribution $p(r_t, \boldsymbol{z}_{1:t}, \boldsymbol{x}_{1:t}, \boldsymbol{\theta}_t)$ by marginalizing over all values of the previous run length r_{t-1}, that is,

$$\begin{aligned} p(r_t, \boldsymbol{z}_{1:t}, \boldsymbol{x}_{1:t}, \boldsymbol{\theta}_t) &= \sum_{r_{t-1}} p(r_t, r_{t-1}, \boldsymbol{z}_{1:t}, \boldsymbol{x}_{1:t}, \boldsymbol{\theta}_t) \\ &= \sum_{r_{t-1}} p(r_t, r_{t-1}, z_t, \boldsymbol{z}_{1:t-1}, \boldsymbol{x}_t, \boldsymbol{x}_{1:t-1}, \boldsymbol{\theta}_t), \end{aligned} \tag{A.1}$$

where we have divided $\boldsymbol{x}_{1:t}$ and $\boldsymbol{z}_{1:t}$ to simplify the derivation. The last term can be rewritten as

$$\begin{aligned} &p(r_t, r_{t-1}, z_t, \boldsymbol{z}_{1:t-1}, \boldsymbol{x}_t, \boldsymbol{x}_{1:t-1}, \boldsymbol{\theta}_t) = \\ &\qquad = p(r_t, z_t, \boldsymbol{x}_t, \boldsymbol{\theta}_t | r_{t-1}, \boldsymbol{z}_{1:t-1}, \boldsymbol{x}_{1:t-1}) p(r_{t-1}, \boldsymbol{z}_{1:t-1}, \boldsymbol{x}_{1:t-1}). \end{aligned} \tag{A.2}$$

where the right-hand side term is

$$p(r_{t-1}, \boldsymbol{z}_{1:t-1}, \boldsymbol{x}_{1:t-1}) = \int p(r_{t-1}, \boldsymbol{z}_{1:t-1}, \boldsymbol{x}_{1:t-1}, \boldsymbol{\theta}_{t-1}) d\boldsymbol{\theta}_{t-1}, \tag{A.3}$$

with $p(r_{t-1}, \boldsymbol{z}_{1:t-1}, \boldsymbol{x}_{1:t-1}, \boldsymbol{\theta}_{t-1})$ being the factorized joint probability distribution at $t-1$.

Since the current run length r_t is only conditioned by its previous value r_{t-1}, Eq. (A.3) can be written down as

$$\begin{aligned} &p(r_t, \boldsymbol{z}_{1:t}, \boldsymbol{x}_{1:t}, \boldsymbol{\theta}_t) = \\ &\qquad = \sum_{r_{t-1}} p(r_t | r_{t-1}) p(z_t, \boldsymbol{x}_t, \boldsymbol{\theta}_t | r_{t-1}, \boldsymbol{z}_{1:t-1}, \boldsymbol{x}_{1:t-1}) p(r_{t-1}, \boldsymbol{z}_{1:t-1}, \boldsymbol{x}_{1:t-1}), \end{aligned} \tag{A.4}$$

where $p(r_t | r_{t-1})$ is the change-point prior. Note that useless conditioned variables have been omitted. At the same time, we may decompose

$$p(z_t, \boldsymbol{x}_t, \boldsymbol{\theta}_t | r_{t-1}, \boldsymbol{z}_{1:t-1}, \boldsymbol{x}_{1:t-1}) = p(\boldsymbol{x}_t | z_t) p(z_t | \boldsymbol{\theta}_t) p(\boldsymbol{\theta}_t | r_{t-1}, \boldsymbol{z}_{1:t-1}), \tag{A.5}$$

where we have taken into account that $\boldsymbol{x}_t$ is only conditioned by its latent representation $\boldsymbol{z}_t$, which is modeled by the likelihood term $p(\boldsymbol{z}_t | \boldsymbol{\theta}_t)$ given the posterior distribution $p(\boldsymbol{\theta}_t | r_{t-1}, \boldsymbol{z}_{1:t-1})$ on the parameters.

The resulting recursive expression is

$$\begin{aligned} &p(r_t, \boldsymbol{z}_{1:t}, \boldsymbol{x}_{1:t}, \boldsymbol{\theta}_t) = \\ &\qquad = \sum_{r_{t-1}} p(r_t | r_{t-1}) p(\boldsymbol{x}_t | z_t) p(z_t | \boldsymbol{\theta}_t) p(\boldsymbol{\theta}_t | r_{t-1}, \boldsymbol{z}_{1:t-1}) p(r_{t-1}, \boldsymbol{z}_{1:t-1}, \boldsymbol{x}_{1:t-1}), \end{aligned} \tag{A.6}$$

and can be calculated sequentially at each time step t.

A.2 Gaussian Likelihood with Missing Data

The Gaussian distribution requires to handle missing or partial observations in this problem. Thus, based on Ghahramani and Jordan (1994), we rewrite the likelihood as follows

$$p(\boldsymbol{x}_i^{\text{real}}|\boldsymbol{\theta}_k) = p(\boldsymbol{x}_i^{o,\text{real}}, \boldsymbol{x}_i^{m,\text{real}}|\boldsymbol{\theta}_k) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{x}_i^o \\ \boldsymbol{x}_i^m\end{bmatrix} \middle| \begin{bmatrix}\mathbf{0}^o \\ \mathbf{0}^m\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_k^{oo} & \boldsymbol{\Sigma}_k^{om} \\ \boldsymbol{\Sigma}_k^{mo} & \boldsymbol{\Sigma}_k^{mm}\end{bmatrix}\right), \tag{A.7}$$

where the blocks of the covariance matrix are given by

$$[\boldsymbol{\Sigma}_k^{oo}]_{t^o,t^{o'}} = g_k(t^o, t^{o'}) + \sigma_t^2 \cdot \mathbb{I}[t^o = t^{o'}], \tag{A.8}$$

$$[\boldsymbol{\Sigma}_k^{mm}]_{t^m,t^{m'}} = g_k(t^m, t^{m'}) + \sigma_t^2 \cdot \mathbb{I}[t^m = t^{m'}] \tag{A.9}$$

$$[\boldsymbol{\Sigma}_k^{mo}]_{t^m,t^{o'}} = g_k(t^m, t^{o'}), \tag{A.10}$$

$$[\boldsymbol{\Sigma}_k^{om}]_{t^o,t^{m'}} = g_k(t^o, t^{m'}), \tag{A.11}$$

with t^o being the time index of an observed variable and t^m the time index of a missing variable.

A.2.1 Expected Complete Heterogeneous Log-Likelihood

The expectation of the complete log-likelihood, which we denote by $\mathcal{Q}$, can be obtained as

$$\begin{aligned}\mathcal{Q} = \mathbb{E}_{\boldsymbol{z}_{1:t},\boldsymbol{x}_{1:t}^m}[\mathcal{L}_{\boldsymbol{z},\boldsymbol{\phi}}] = &\sum_{i=1}^{t}\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{z}_i}\Big[\mathbb{I}\{z_i = k\}\Big]\log\pi_k \\ &+ \sum_{i=1}^{t}\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{z}_i}\Big[\mathbb{I}\{z_i = k\}\Big]\mathbb{E}_{\boldsymbol{x}^m}\Big[\log p(\boldsymbol{x}_i^o, \boldsymbol{x}_i^m|\boldsymbol{\theta}_k)\Big].\end{aligned} \tag{A.12}$$

Substituting $p(\boldsymbol{x}_i^o, \boldsymbol{x}_i^m|\boldsymbol{\theta}_k)$ by the Bernoulli-Gaussian mixture, we obtain

$$\begin{aligned}\mathcal{Q} = &\sum_{i=1}^{t}\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{z}}\Big[\mathbb{I}\{z_i = k\}\Big]\log\pi_k + \sum_{i=1}^{t}\sum_{k=1}^{K}\mathbb{E}_{\boldsymbol{z}}\Big[\mathbb{I}\{z_i = k\}\Big]\Big\{-\frac{D}{2}\log(2\pi) \\ &- \frac{1}{2}\log\left(|\boldsymbol{\Sigma}_k|\right) - \frac{1}{2}\left(\boldsymbol{x}_i^{\text{real},o}\right)^\top\left(\boldsymbol{\Sigma}_k^{oo}\right)^{-1}\boldsymbol{x}_i^{\text{real},o} \\ &- \frac{1}{2}\mathbb{E}_{\boldsymbol{x}^{\text{real},m}}\left[\boldsymbol{x}_i^{\text{real},m}\right]^\top\left(\boldsymbol{\Sigma}_k^{mo}\right)^{-1}\boldsymbol{x}_i^{\text{real},o} \\ &- \frac{1}{2}\left(\boldsymbol{x}_i^{\text{real},o}\right)^\top\left(\boldsymbol{\Sigma}_k^{om}\right)^{-1}\mathbb{E}_{\boldsymbol{x}^{\text{real},m}}\left[\boldsymbol{x}_i^{\text{real},m}\right] \\ &- \frac{1}{2}\mathbb{E}_{\boldsymbol{x}^{\text{real},m}}\left[\boldsymbol{x}_i^{\text{real},m}\right]^\top\left(\boldsymbol{\Sigma}_k^{mm}\right)^{-1}\mathbb{E}_{\boldsymbol{x}^{\text{real},m}}\left[\boldsymbol{x}_i^{\text{real},m}\right] \\ &+ \frac{1}{2}\text{tr}(\boldsymbol{\Sigma}_k^{-1}\text{Cov}(\boldsymbol{x}_i^{\text{real},m})) + \mathbb{E}_{\boldsymbol{x}^{\text{bin},m}}\left[\boldsymbol{x}_i^{\text{bin},m}\right]\log\boldsymbol{\mu}_k^m \\ &+ \left(1 - \mathbb{E}_{\boldsymbol{x}^{\text{bin},m}}\left[\boldsymbol{x}_i^{\text{bin},m}\right]\right)\log(1 - \boldsymbol{\mu}_k^m) \\ &+ \boldsymbol{x}_i^{\text{bin},o}\log\boldsymbol{\mu}_k^o + \left(1 - \boldsymbol{x}_i^{\text{bin},o}\right)\log(1 - \boldsymbol{\mu}_k^o)\Big\}.\end{aligned} \tag{A.13}$$

A.2.2 Derivatives of the Heterogeneous Log-Likelihood

Let us denote the set of hyperparameters of the periodic non-stationary kernel $g_k(t,t')$ as $\boldsymbol{\psi}_k = [\ell_k, \mathbf{a}_k^\top, \mathbf{b}_k^\top]^\top$. Additionally, we refer to all variables $\widetilde{\boldsymbol{x}}$ as the ones whose missing values $\boldsymbol{x}^m$ have been replaced by the expected ones at each E-step, similarly to the convention adopted in Ghahramani and Jordan (1994). The derivatives of $\mathcal{Q}$ w.r.t. $\boldsymbol{\psi}_k$ and $\boldsymbol{\sigma}$ are respectively

$$\begin{aligned}
\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\psi}_k} &= \frac{\partial}{\partial \boldsymbol{\psi}_k}\left[-\frac{1}{2}\sum_{i=1}^{t}\sum_{k=1}^{K} r_{ik}^o \log(|\boldsymbol{\Sigma}_k|) - \frac{1}{2}\sum_{i=1}^{t}\sum_{k=1}^{K} r_{ik}^o \widetilde{\boldsymbol{x}}_i^{\text{real}\top}\boldsymbol{\Sigma}_k^{-1}\widetilde{\boldsymbol{x}}_i^{\text{real}}\right] \\
&- \frac{\partial}{\partial \boldsymbol{\psi}_k}\left[\frac{1}{2}\sum_{i=1}^{t}\sum_{k=1}^{K} r_{ik}^o \text{tr}(\boldsymbol{\Sigma}_k^{-1}\text{Cov}(\boldsymbol{x}_i^m))\right] \\
&= \frac{1}{2}\sum_{i=1}^{t} r_{ik}^o \text{tr}\left(\left(\boldsymbol{\alpha}_{ik}\boldsymbol{\alpha}_{ik}^\top - \boldsymbol{\Sigma}_k^{-1}\right)\frac{\partial \boldsymbol{\Sigma}_k}{\partial \boldsymbol{\psi}_k}\right) + \frac{1}{2}\sum_{i=1}^{N} r_{ik}^o \text{tr}\left(\left(\boldsymbol{\Sigma}_k^{-1}\mathbf{A}_k^{\text{old}}\boldsymbol{\Sigma}_k^{-1}\right)\frac{\partial \boldsymbol{\Sigma}_k}{\partial \boldsymbol{\psi}_k}\right),
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial \mathcal{Q}}{\partial \boldsymbol{\sigma}} &= \frac{\partial}{\partial \boldsymbol{\sigma}}\left[-\frac{1}{2}\sum_{i=1}^{t}\sum_{k=1}^{K} r_{ik}^o \log(|\boldsymbol{\Sigma}_k|) - \frac{1}{2}\sum_{i=1}^{t}\sum_{k=1}^{K} r_{ik}^o \widetilde{\mathbf{x}}_i^{\text{real}\top}\boldsymbol{\Sigma}_k^{-1}\widetilde{\boldsymbol{x}}_i^{\text{real}}\right] \\
&- \frac{\partial}{\partial \boldsymbol{\sigma}}\left[\frac{1}{2}\sum_{i=1}^{t}\sum_{k=1}^{K} r_{ik}^o \text{tr}\left(\boldsymbol{\Sigma}_k^{-1}\text{Cov}(\boldsymbol{x}_i^m)\right)\right] \\
&= \frac{1}{2}\sum_{i=1}^{t}\sum_{k=1}^{K} r_{ik}^o \text{tr}\left(\left(\boldsymbol{\alpha}_{ik}\boldsymbol{\alpha}_{ik}^\top - \boldsymbol{\Sigma}_k^{-1}\right)\frac{\partial \boldsymbol{\Sigma}_k}{\partial \boldsymbol{\sigma}}\right) \\
&+ \frac{1}{2}\sum_{i=1}^{t}\sum_{k=1}^{K} r_{ik}^o \text{tr}\left(\left(\boldsymbol{\Sigma}_k^{-1}\mathbf{A}_k^{\text{old}}\boldsymbol{\Sigma}_k^{-1}\right)\frac{\partial \boldsymbol{\Sigma}_k}{\partial \boldsymbol{\sigma}}\right),
\end{aligned} \tag{A.14}$$

where $\boldsymbol{\alpha}_{ik} = \boldsymbol{\Sigma}_k^{-1}\widetilde{\boldsymbol{x}}_i$, with

$$\widetilde{\boldsymbol{x}}_i^{\text{real}} = \begin{bmatrix} \boldsymbol{x}_i^{\text{real},o} \\ \mathbb{E}_{\boldsymbol{x}^{\text{real},m}}[\boldsymbol{x}_i^{\text{real},m}] \end{bmatrix}, \qquad \mathbf{A}_k^{\text{old}} = \begin{bmatrix} \mathbf{0}^{oo} & \mathbf{0}^{om} \\ \mathbf{0}^{mo} & \text{Cov}^{\text{old}}(\boldsymbol{x}_k^m) \end{bmatrix}. \tag{A.15}$$

and $\text{Cov}^{\text{old}}(\cdot)$ is the covariance sub-matrix of the given missing dimensions. Here, we have used that

$$\frac{\partial \boldsymbol{\Sigma}_k}{\partial \boldsymbol{\psi}_k} = \frac{\partial \mathbf{K}_k}{\partial \boldsymbol{\psi}_k}, \tag{A.16}$$

and

$$\frac{\partial \boldsymbol{\Sigma}_k}{\partial \boldsymbol{\sigma}} = \frac{\partial \mathbf{D}}{\partial \boldsymbol{\sigma}}. \tag{A.17}$$

A.2.3 Derivatives of the Periodic Non-stationary Kernel

In this appendix, we present the derivatives w. r. t. ℓ_k, $\mathbf{a}_k$ and $\mathbf{b}_k$, which are given by

$$
\begin{aligned}
\frac{\partial g_k(t,t')}{\partial \ell_k} &= s_k(t)^2 s_k(t')^2 \exp\left(-\frac{2\sin^2(\pi(t-t')/T)}{\ell_k^2}\right)\left(\frac{4\sin^2(\pi(t-t')/T)}{\ell_k^3}\right) \\
\frac{\partial g_k(t,t')}{\partial a_{0k}} &= s_k(t) s_k(t') \exp\left(-\frac{2\sin^2(\pi(t-t')/T)}{\ell_k^2}\right)(s_k(t)+s_k(t')) \\
\frac{\partial g_k(t,t')}{\partial a_{nk}} &= 2\exp\left(-\frac{2\sin^2(\pi(t-t')/T)}{\ell_k^2}\right) s_k(t) s_k(t') \\
&\quad \times \left(\cos\left(\frac{2\pi n}{T}t\right) s_k(t') + \cos\left(\frac{2\pi n}{T}t'\right) s_k(t)\right) \\
\frac{\partial g_k(t,t')}{\partial b_{nk}} &= 2\exp\left(-\frac{2\sin^2(\pi(t-t')/T)}{\ell_k^2}\right) s_k(t) s_k(t') \\
&\quad \times \left(\sin\left(\frac{2\pi n}{T}t\right) s_k(t') + \sin\left(\frac{2\pi n}{T}t'\right) s_k(t)\right)
\end{aligned}
$$

APPENDIX B

HETEROGENEOUS MOGP DERIVATIONS

B.1 Derivation of Heterogeneous Multi-output Lower Bounds

$$
\begin{aligned}
\log p(\boldsymbol{y}) &= \log \int p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u})d\boldsymbol{f}d\boldsymbol{u} \\
&= \log \int q(\boldsymbol{f},\boldsymbol{u})\frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u})}{q(\boldsymbol{f},\boldsymbol{u})}d\boldsymbol{f}d\boldsymbol{u} \qquad \text{(B.1)} \\
\mathcal{L} &= \int q(\boldsymbol{f},\boldsymbol{u})\log\frac{p(\boldsymbol{y}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{u})p(\boldsymbol{u})}{q(\boldsymbol{f},\boldsymbol{u})}d\boldsymbol{f}d\boldsymbol{u} \\
&= \int \prod_{d=1}^{D}\prod_{j=1}^{J_d} p(\boldsymbol{f}_{d,j}|\boldsymbol{u})q(\boldsymbol{u})\log\frac{p(\boldsymbol{y}|\boldsymbol{f})\prod_{d=1}^{D}\prod_{j=1}^{J_d}p(\boldsymbol{f}_{d,j}|\boldsymbol{u})p(\boldsymbol{u})}{\prod_{d=1}^{D}\prod_{j=1}^{J_d}p(\boldsymbol{f}_{d,j}|\boldsymbol{u})q(\boldsymbol{u})}d\boldsymbol{f}d\boldsymbol{u} \\
&= \int \prod_{d=1}^{D}\prod_{j=1}^{J_d} p(\boldsymbol{f}_{d,j}|\boldsymbol{u})q(\boldsymbol{u})\log p(\boldsymbol{y}|\boldsymbol{f})d\boldsymbol{f}d\boldsymbol{u} - \sum_{q=1}^{Q}\mathrm{KL}\big(q(\boldsymbol{u}_q)||p(\boldsymbol{u}_q)\big) \\
&= \int \prod_{d=1}^{D}\prod_{j=1}^{J_d} p(\boldsymbol{f}_{d,j}|\boldsymbol{u})q(\boldsymbol{u})d\boldsymbol{u}\log p(\boldsymbol{y}|\boldsymbol{f})d\boldsymbol{f} - \sum_{q=1}^{Q}\mathrm{KL}\big(q(\boldsymbol{u}_q)||p(\boldsymbol{u}_q)\big) \\
&= \int q(\boldsymbol{f})\log p(\boldsymbol{y}|\boldsymbol{f})d\boldsymbol{f} - \sum_{q=1}^{Q}\mathrm{KL}\big(q(\boldsymbol{u}_q)||p(\boldsymbol{u}_q)\big) \\
&= \int q(\boldsymbol{f})\sum_{d=1}^{D}\log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)d\boldsymbol{f} - \sum_{q=1}^{Q}\mathrm{KL}\big(q(\boldsymbol{u}_q)||p(\boldsymbol{u}_q)\big) \\
&= \sum_{d=1}^{D}\int q(\widetilde{\boldsymbol{f}}_d)\log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)d\widetilde{\boldsymbol{f}}_d - \sum_{q=1}^{Q}\mathrm{KL}\big(q(\mathbf{u}_q)||p(\mathbf{u}_q)\big) \\
&= \sum_{d=1}^{D}\mathbb{E}_{q(\widetilde{\boldsymbol{f}}_d)}\big[\log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)\big] - \sum_{q=1}^{Q}\mathrm{KL}\big(q(\boldsymbol{u}_q)||p(\boldsymbol{u}_q)\big). \qquad \text{(B.2)}
\end{aligned}
$$

B.2 Gradients w.r.t. $q(\mathbf{u})$

First, the bound derivatives w.r.t. $\boldsymbol{\mu}_{\boldsymbol{u}_q}$ are

$$
\frac{\partial}{\partial\boldsymbol{\mu}_{\boldsymbol{u}_q}}\mathcal{L} = \sum_{d=1}^{D}\underbrace{\frac{\partial}{\partial\boldsymbol{\mu}_{\boldsymbol{u}_q}}\mathbb{E}_{q(\widetilde{\boldsymbol{f}}_d)}\big\{\log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)\big\}}_{\text{VE part}} - \underbrace{\frac{\partial}{\partial\boldsymbol{\mu}_{\boldsymbol{u}_q}}\mathrm{KL}\big(q(\boldsymbol{u}_q)||p(\boldsymbol{u}_q)\big)}_{\text{KL part}}, \qquad \text{(B.3)}
$$

where the KL part w.r.t. $\boldsymbol{\mu}_{\boldsymbol{u}_q}$ is

$$\frac{\partial}{\partial \boldsymbol{\mu}_{\boldsymbol{u}_q}} \mathrm{KL}\big(q(\boldsymbol{u}_q)|p(\boldsymbol{u}_q)\big) = \mathbf{K}_q^{-1}\boldsymbol{\mu}_{\boldsymbol{u}_q}, \tag{B.4}$$

and the VE part w.r.t. $\boldsymbol{\mu}_{\boldsymbol{u}_q}$ yields

$$\frac{\partial}{\partial \boldsymbol{\mu}_{\boldsymbol{u}_q}} \mathbb{E}_{q(\widetilde{\boldsymbol{f}}_d)}\big[\log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)\big] = \mathbb{E}_{q(\mathbf{F}^t)}\big[\underbrace{\frac{\partial}{\partial \widetilde{\boldsymbol{f}}_d}\log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)}_{\text{See Likelihoods}}\big]\frac{\partial \widetilde{\boldsymbol{m}}_d}{\partial \boldsymbol{\mu}_{\boldsymbol{u}_q}}. \tag{B.5}$$

Secondly, the bound derivatives w.r.t. $\mathbf{S}_{\mathbf{u}_q}$ are

$$\frac{\partial}{\partial \mathbf{S}_{\boldsymbol{u}_q}}\mathcal{L} = \sum_{d=1}^{D}\underbrace{\frac{\partial}{\partial \mathbf{S}_{\boldsymbol{u}_q}}\mathbb{E}_{q(\widetilde{\boldsymbol{f}}_d)}\big[\log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)\big]}_{\text{VE part}} - \underbrace{\frac{\partial}{\partial \mathbf{S}_{\boldsymbol{u}_q}}\mathrm{KL}\big(q(\boldsymbol{u}_q)||p(\boldsymbol{u}_q)\big)}_{\text{KL part}}, \tag{B.6}$$

where the KL part w.r.t. $\mathbf{S}_{\boldsymbol{u}_q}$ is

$$\frac{\partial}{\partial \mathbf{S}_{\boldsymbol{u}_q}}\mathrm{KL}\big(q(\boldsymbol{u}_q)||p(\boldsymbol{u}_q)\big) = \mathbf{K}_q^{-1} - \frac{1}{2}\mathrm{diag}(\mathbf{K}_q^{-1}) - \frac{1}{2}\mathbf{S}_{\boldsymbol{u}_q}^{-1}, \tag{B.7}$$

and the VE part w.r.t. $\mathbf{S}_{\boldsymbol{u}_q}$ yields

$$\frac{\partial}{\partial \mathbf{S}_{\boldsymbol{u}_q}}\mathbb{E}_{q(\widetilde{\boldsymbol{f}}_d)}\big[\log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)\big] = \frac{1}{2}\mathbb{E}_{q(\widetilde{\boldsymbol{f}}_d)}\big[\underbrace{\frac{\partial^2}{\partial \widetilde{\boldsymbol{f}}_d^2}\log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)}_{\text{See Likelihoods}}\big]\frac{\partial \widetilde{\boldsymbol{v}}_d}{\partial \mathbf{S}_{\boldsymbol{u}_q}}. \tag{B.8}$$

where $\widetilde{\boldsymbol{m}}_d$ and $\widetilde{\boldsymbol{v}}_d$ are the corresponding mean and variance of the variational distribution $q(\widetilde{\boldsymbol{f}}_d)$. Each one of the variational expectations on the functional derivatives is different for a given heterogeneous likelihood (see below). The gradients identities in (B.5) and (B.8) are similar to the ones used in Opper and Archambeau (2009); Hensman et al. (2013a); Saul et al. (2016). This means to use

$$\frac{\partial}{\partial \sigma}\mathbb{E}_{\mathcal{N}(x|\mu,\sigma^2)}\big[f(x)\big] = \mathbb{E}_{\mathcal{N}(x|\mu,\sigma^2)}\big[\frac{\partial}{\partial x}f(x)\big], \tag{B.9}$$

$$\frac{\partial}{\partial \mu}\mathbb{E}_{\mathcal{N}(x|\mu,\sigma^2)}\big[f(x)\big] = \frac{1}{2}\mathbb{E}_{\mathcal{N}(x|\mu,\sigma^2)}\big[\frac{\partial^2}{\partial x^2}f(x)\big]. \tag{B.10}$$

B.3 Gradients w.r.t hyperparameters

Applying the *chain-rule* and assuming the matrix derivatives $\frac{\partial \mathbf{A}}{\partial \boldsymbol{\theta}}$ and $\frac{\partial \mathbf{A}}{\partial \mathbf{Z}}$ given for any arbitrary matrix $\mathbf{A}$ dependent on the hyperparameters, we must compute the following gradients:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{K}_q}, \frac{\partial \mathcal{L}}{\partial \mathbf{K}_{\boldsymbol{f}_{d,j}\boldsymbol{u}_q}} \text{ and } \frac{\partial \mathcal{L}}{\partial \mathrm{diag}(\mathbf{K}_{\boldsymbol{f}_{d,j}\boldsymbol{f}_{d,j}})}. \tag{B.11}$$

In this section, we denote $\mathbf{K}_{dq} = \mathbf{K}_{\boldsymbol{f}_{d,j}\boldsymbol{u}_q}$ and $\mathbf{K}_{\mathrm{diag}} = \mathrm{diag}(\mathbf{K}_{\boldsymbol{f}_{d,j}\boldsymbol{f}_{d,j}})$ for simplicity in the following expressions. Then,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{K}_q} = \sum_{d=1}^{D} \frac{\partial}{\partial \mathbf{K}_q} \mathbb{E}_{q(\widetilde{\boldsymbol{f}}_d)}\big[\log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)\big] - \sum_{q=1}^{Q} \frac{\partial}{\partial \mathbf{K}_q} \mathrm{KL}\big(q(\boldsymbol{u}_q)||p(\boldsymbol{u}_q)\big), \tag{B.12}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{K}_{dq}} = \sum_{d=1}^{D} \sum_{j=1}^{J_d} \frac{\partial}{\partial \mathbf{K}_{dq}} \mathbb{E}_{q(\widetilde{\boldsymbol{f}}_d)}\big[\log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)\big], \tag{B.13}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{K}_{\mathrm{diag}}} = \sum_{d=1}^{D} \sum_{j=1}^{J_d} \frac{\partial}{\partial \mathbf{K}_{\mathrm{diag}}} \mathbb{E}_{q(\widetilde{\boldsymbol{f}}_d)}\big[\log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)\big]; \tag{B.14}$$

where

$$\frac{\partial}{\partial \mathbf{K}_q} \mathrm{KL}\big(q(\boldsymbol{u}_q)||p(\boldsymbol{u}_q)\big) = \frac{1}{2}\Bigg(- (\mathbf{K}_q^{-1}\mathbf{S}_{\boldsymbol{u}_q}\mathbf{K}_q^{-1})^\top - (\mathbf{K}_q^{-1})^\top \boldsymbol{\mu}_{\boldsymbol{u}_q}\boldsymbol{\mu}_{\boldsymbol{u}_q}^\top (\mathbf{K}_q^{-1})^\top + (\mathbf{K}_q^{-1})^\top \Bigg), \tag{B.15}$$

and

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{K}_q} \mathbb{E}_{q(\widetilde{\boldsymbol{f}}_d)}\big[\log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)\big] &= \frac{\partial}{\partial \widetilde{\boldsymbol{m}}_d} \mathbb{E}_{q(\widetilde{\boldsymbol{f}}_d)}\big[\log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)\big] \frac{\partial \widetilde{\boldsymbol{m}}_d}{\partial \mathbf{K}_q} \\
&+ \frac{\partial}{\partial \widetilde{\mathbf{v}}_d} \mathbb{E}_{q(\widetilde{\boldsymbol{f}}_d)}\big[\log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)\big] \frac{\partial \widetilde{\boldsymbol{v}}_d}{\partial \mathbf{K}_q},
\end{aligned} \tag{B.16}$$

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{K}_{dq}} \mathbb{E}_{q(\widetilde{\boldsymbol{f}}_d)}\big[\log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)\big] &= \frac{\partial}{\partial \widetilde{\boldsymbol{m}}_d} \mathbb{E}_{q(\widetilde{\boldsymbol{f}}_d)}\big[\log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)\big] \frac{\partial \widetilde{\boldsymbol{m}}_d}{\partial \mathbf{K}_{dq}} \\
&+ \frac{\partial}{\partial \widetilde{\boldsymbol{v}}_d} \mathbb{E}_{q(\widetilde{\boldsymbol{f}}_d)}\big[\log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)\big] \frac{\partial \widetilde{\boldsymbol{v}}_d}{\partial \mathbf{K}_{dq}},
\end{aligned} \tag{B.17}$$

$$\frac{\partial}{\partial \mathbf{K}_{\mathrm{diag}}} \mathbb{E}_{q(\widetilde{\boldsymbol{f}}_d)}\big[\log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)\big] = \frac{\partial}{\partial \widetilde{\boldsymbol{v}}_d} \mathbb{E}_{q(\widetilde{\boldsymbol{f}}_d)}\big[\log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)\big] \frac{\partial \widetilde{\boldsymbol{v}}_d}{\partial \mathbf{K}_{\mathrm{diag}}}.$$

B.4 Likelihoods and link functions

To include any new distribution, we must derive the following expressions for each heterogeneous likelihood $p(\mathbf{y}_d|\widetilde{\mathbf{f}}_d)$:

1. Log-Likelihood function $\log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)$ for VE and the predictive distribution.
2. First order derivatives $\frac{\partial}{\partial \widetilde{\boldsymbol{f}}_d} \log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)$ for VE in gradients.
3. Second order derivatives $\frac{\partial^2}{\partial \widetilde{\boldsymbol{f}}_d^2} \log p(\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d)$ for VE in gradients.
4. Mean $\mathbb{E}\big[\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d\big]$ and variance $\mathrm{var}\big[\boldsymbol{y}_d|\widetilde{\boldsymbol{f}}_d\big]$ for predictive point-estimates.

B.4.1 Heterogeneous likelihood syntaxes

In our code, we implemented a simple manner to define the heterogeneous likelihood for combinations of an arbitrary number of likelihood functions. The assignment of LPFs to parameters is done automatically. Some examples are given below:

- `likelihood_list = [Gaussian(), Gaussian(sigma=0.5), Exponential()]`

Table B.1: List of the used linking transformations between latent parameter functions (LPFs) $\boldsymbol{f}$ and the heterogeneous likelihoods. Note that many other valid mappings between parameters and LPFs are allowed.

LIKELIHOOD	LINKED PARAMETERS	Number of LPFs $\boldsymbol{f}$
Gaussian	$\mu(\boldsymbol{x}) = \boldsymbol{f}$, $\sigma(\boldsymbol{x})$	1
Heteroscedastic Gaussian	$\mu(\boldsymbol{x}) = \boldsymbol{f}_1$, $\sigma(\boldsymbol{x}) = \exp(\boldsymbol{f}_2)$	2
Bernoulli	$\rho(\boldsymbol{x}) = \frac{\exp(\boldsymbol{f})}{1+\exp(\boldsymbol{f})}$	1
Categorical	$\rho_k(\boldsymbol{x}) = \frac{\exp(\boldsymbol{f}_k)}{1+\sum_{k'=1}^{K-1}\exp(\boldsymbol{f}_{k'})}$	K-1
Exponential	$\beta(\boldsymbol{x}) = \exp(-\boldsymbol{f})$	1
Poisson	$\lambda(\boldsymbol{x}) = \exp(\boldsymbol{f})$	1
Gamma	$a(\boldsymbol{x}) = \exp(\boldsymbol{f}_1), b(\boldsymbol{x}) = \exp(\boldsymbol{f}_2)$	2
Beta	$a(\boldsymbol{x}) = \exp(\boldsymbol{f}_1), b(\boldsymbol{x}) = \exp(\boldsymbol{f}_2)$	2

- `likelihood_list = [HetGaussian(), Bernoulli(), Categorical(K=3)]`
- `likelihood_list = [Gamma(), Categorical(K=5)]`

APPENDIX C

RECYCLABLE GAUSSIAN PROCESSES

C.1 Detailed Derivation of the Lower Ensemble Bound

The construction of ensemble variational bounds from recyclable GP models is based on the idea of *augmenting* the marginal likelihood to be conditioned on the infinite-dimensional GP function f_∞. Notice that f_∞ contains all the function values taken by $f(\cdot)$ over the input-space $\mathbb{R}^p$, including the input targets $\{\boldsymbol{x}_i\}_{i=1}^N$, the local *inducing-inputs* $\{\boldsymbol{Z}_k\}_{k=1}^K$ and the global ones $\boldsymbol{Z}_*$. Having K partitions of the dataset $\mathcal{D}$ with their corresponding outputs $\boldsymbol{y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_K\}$, we begin by augmenting the marginal log-likelihood as

$$\log p(\boldsymbol{y}) = \log p(\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_K) = \log \int p(\boldsymbol{y}, f_\infty) df_\infty, \tag{C.1}$$

that factorises according to

$$\log \int p(\boldsymbol{y}, f_\infty) df_\infty = \log \int p(\boldsymbol{y}|f_\infty) p(f_\infty) df_\infty, \tag{C.2}$$

where $p(\boldsymbol{y}|f_\infty)$ is the *augmented* likelihood term of all the output targets of interest and $p(f_\infty)$ the GP prior over the infinite amount of points in the input-space $\mathbb{R}^p$. This last distribution takes the form of an infinite-dimensional Gaussian, that we avoid to evaluate explicitly in the equations. To build the lower bound on the log-marginal likelihood, we first introduce the global variational distribution $q(\boldsymbol{u}_*) = \mathcal{N}(\boldsymbol{u}_*|\boldsymbol{\mu}_*, \boldsymbol{S}_*)$ into the equation,

$$\begin{aligned}\log p(\boldsymbol{y}) = \log \int p(\boldsymbol{y}|f_\infty) p(f_\infty) df_\infty &= \log \int \frac{q(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} p(\boldsymbol{y}|f_\infty) p(f_\infty) df_\infty \\ &= \log \iint \frac{q(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} p(\boldsymbol{y}|f_\infty) p(f_{\infty \neq \boldsymbol{u}_*}|\boldsymbol{u}_*) p(\boldsymbol{u}_*) df_{\infty \neq \boldsymbol{u}_*} d\boldsymbol{u}_*. \end{aligned} \tag{C.3}$$

Notice that the differentials df_∞ have been splitted into $df_{\infty \neq \boldsymbol{u}_*} d\boldsymbol{u}_*$, and at the same time, we applied properties of Gaussian conditionals in the GP prior to rewrite $p(f_\infty)$ as $p(f_{\infty \neq \boldsymbol{u}_*}|\boldsymbol{u}_*)p(\boldsymbol{u}_*)$. When the target variables $\boldsymbol{u}_*$ are explicit in the expression, our second step is the application of the Jensen inequality twice as it is done in the reparameterisation of (Gal et al., 2014), that is

$$\log p(\boldsymbol{y}) = \log \iint \frac{q(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} p(\boldsymbol{y}|f_\infty) p(f_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*) p(\boldsymbol{u}_*) df_{\infty\neq\boldsymbol{u}_*} d\boldsymbol{u}_*$$
$$= \log \iint q(\boldsymbol{u}_*) p(f_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*) p(\boldsymbol{y}|f_\infty) \frac{p(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} df_{\infty\neq\boldsymbol{u}_*} d\boldsymbol{u}_*$$
$$= \log \left(\mathbb{E}_{q(\boldsymbol{u}_*)} \left[\mathbb{E}_{p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*)} \left[p(\boldsymbol{y}|f_\infty) \frac{p(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} \right] \right] \right)$$
$$\geq \mathbb{E}_{q(\boldsymbol{u}_*)} \left[\log \left(\mathbb{E}_{p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*)} \left[p(\boldsymbol{y}|f_\infty) \frac{p(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} \right] \right) \right]$$
$$\geq \mathbb{E}_{q(\boldsymbol{u}_*)} \left[\mathbb{E}_{p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*)} \left[\log \left(p(\boldsymbol{y}|f_\infty) \frac{p(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} \right) \right] \right] = \mathcal{L}_\mathcal{E}. \quad \text{(C.4)}$$

Then, if we have Eq. (C.4), which is the first version of our ensemble lower bound $\mathcal{L}_\mathcal{E}$, we can use the augmented likelihood term $p(\boldsymbol{y}|f_\infty)$ to introduce the local approximations to f instead of revisiting the data. This is,

$$\mathcal{L}_\mathcal{E} = \mathbb{E}_{q(\boldsymbol{u}_*)} \left[\mathbb{E}_{p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*)} \left[\log p(\boldsymbol{y}|f_\infty) + \log \left(\frac{p(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} \right) \right] \right]$$
$$= \mathbb{E}_{q(\boldsymbol{u}_*)} \left[\mathbb{E}_{p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*)} \left[\log p(\boldsymbol{y}|f_\infty) \right] - \log \left(\frac{q(\boldsymbol{u}_*)}{p(\boldsymbol{u}_*)} \right) \right]$$
$$= \mathbb{E}_{q(\boldsymbol{u}_*)} \left[\mathbb{E}_{p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*)} \left[\sum_{k=1}^{K} \log p(\boldsymbol{y}_k|f_\infty) \right] - \log \left(\frac{q(\boldsymbol{u}_*)}{p(\boldsymbol{u}_*)} \right) \right]$$
$$= \mathbb{E}_{q(\boldsymbol{u}_*)} \left[\sum_{k=1}^{K} \mathbb{E}_{p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*)} \left[\log p(\boldsymbol{y}_k|f_\infty) \right] - \log \left(\frac{q(\boldsymbol{u}_*)}{p(\boldsymbol{u}_*)} \right) \right], \quad \text{(C.5)}$$

where the log-ratio $q(\boldsymbol{u}_*)/p(\boldsymbol{u}_*)$ acts as a constant to the second expectation $\mathbb{E}_{p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*)}[\cdot]$ and we applied *conditional independence* (CI) among all the output partitions given the latent function f_∞. That is, we introduced $p(\boldsymbol{y}|f_\infty) = \prod_{k=1}^{K} p(\boldsymbol{y}_k|f_\infty)$ to factorise the expectation term in Eq. (C.5) across the K tasks.

Under the approximation of $p(\boldsymbol{y}_k|f_\infty)$ obtained by inverting the Bayes theorem, we use $p(\boldsymbol{y}_k|f_\infty) \approx q_k(f_\infty)/p_k(f_\infty)$ to introduce the local posterior distributions $q_k(\cdot)$ and priors $p_k(\cdot)$ in the bound $\mathcal{L}_\mathcal{E}$. This leads to

$$\mathcal{L}_\mathcal{E} = \mathbb{E}_{q(\boldsymbol{u}_*)} \left[\sum_{k=1}^{K} \mathbb{E}_{p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*)} \left[\log p(\boldsymbol{y}_k|f_\infty) \right] - \log \left(\frac{q(\boldsymbol{u}_*)}{p(\boldsymbol{u}_*)} \right) \right]$$
$$\approx \mathbb{E}_{q(\boldsymbol{u}_*)} \left[\sum_{k=1}^{K} \mathbb{E}_{p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*)} \left[\log \left(\frac{q_k(f_\infty)}{p_k(f_\infty)} \right) \right] - \log \left(\frac{q(\boldsymbol{u}_*)}{p(\boldsymbol{u}_*)} \right) \right]$$
$$= \mathbb{E}_{q(\boldsymbol{u}_*)} \left[\sum_{k=1}^{K} \mathbb{E}_{p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*)} \left[\log \left(\frac{q_k(\boldsymbol{u}_k)}{p_k(\boldsymbol{u}_k)} \right) \right] - \log \left(\frac{q(\boldsymbol{u}_*)}{p(\boldsymbol{u}_*)} \right) \right], \quad \text{(C.6)}$$

where we now have the explicit local distributions $q_k(\boldsymbol{u}_k)$ and $p_k(\boldsymbol{u}_k)$ on the subsets of inducing-inputs $\{\boldsymbol{Z}_k\}_{k=1}^{K}$. The cancellation of conditionals is a result of the variational factorization (Titsias, 2009a). Looking to the last version of the bound in Eq. (C.6), there is

still one point that maintains the infinite-dimensionality, the conditional prior $p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*)$ and its corresponding expectation term $\mathbb{E}_{p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*)}[\cdot]$. To adapt it to the local inducing variables $\boldsymbol{u}_k$, we apply the following simplification to each k-th integral in Eq. (C.6) based in the properties of Gaussian marginals (see next sections),

$$\begin{aligned}\mathbb{E}_{p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*)}\left[\log\left(\frac{q_k(\boldsymbol{u}_k)}{p_k(\boldsymbol{u}_k)}\right)\right] &= \int p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*)\log\left(\frac{q_k(\boldsymbol{u}_k)}{p_k(\boldsymbol{u}_k)}\right)d\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}\\ &= \iint p(\boldsymbol{f}_{\infty\neq\{\boldsymbol{u}_*,\boldsymbol{u}_k\}},\boldsymbol{u}_k|\boldsymbol{u}_*)\log\left(\frac{q_k(\boldsymbol{u}_k)}{p_k(\boldsymbol{u}_k)}\right)d\boldsymbol{f}_{\infty\neq\{\boldsymbol{u}_*,\boldsymbol{u}_k\}}d\boldsymbol{u}_k\\ &= \int p(\boldsymbol{u}_k|\boldsymbol{u}_*)\log\left(\frac{q_k(\boldsymbol{u}_k)}{p_k(\boldsymbol{u}_k)}\right)d\boldsymbol{u}_k = \mathbb{E}_{p(\boldsymbol{u}_k|\boldsymbol{u}_*)}\left[\log\left(\frac{q_k(\boldsymbol{u}_k)}{p_k(\boldsymbol{u}_k)}\right)\right]. \quad \text{(C.7)}\end{aligned}$$

This is the expectation that we plug in the final version of the bound, to obtain

$$\begin{aligned}\mathcal{L}_{\mathcal{E}} &= \mathbb{E}_{q(\boldsymbol{u}_*)}\left[\sum_{k=1}^{K}\mathbb{E}_{p(\boldsymbol{u}_k|\boldsymbol{u}_*)}\left[\log\left(\frac{q_k(\boldsymbol{u}_k)}{p_k(\boldsymbol{u}_k)}\right)\right]-\log\left(\frac{q(\boldsymbol{u}_*)}{p(\boldsymbol{u}_*)}\right)\right]\\ &= \sum_{k=1}^{K}\mathbb{E}_{q(\boldsymbol{u}_*)}\left[\mathbb{E}_{p(\boldsymbol{u}_k|\boldsymbol{u}_*)}\left[\log\left(\frac{q_k(\boldsymbol{u}_k)}{p_k(\boldsymbol{u}_k)}\right)\right]\right]-\mathbb{E}_{q(\boldsymbol{u}_*)}\left[\log\left(\frac{q(\boldsymbol{u}_*)}{p(\boldsymbol{u}_*)}\right)\right]\\ &= \sum_{k=1}^{K}\mathbb{E}_{q(\boldsymbol{u}_*)}\left[\mathbb{E}_{p(\boldsymbol{u}_k|\boldsymbol{u}_*)}\left[\log\left(\frac{q_k(\boldsymbol{u}_k)}{p_k(\boldsymbol{u}_k)}\right)\right]\right]-\mathrm{KL}\left[q(\boldsymbol{u}_*)||p(\boldsymbol{u}_*)\right]\\ &= \sum_{k=1}^{K}\mathbb{E}_{q_{\mathcal{C}}(\boldsymbol{u}_k)}\left[\log q_k(\boldsymbol{u}_k)-\log p_k(\boldsymbol{u}_k)\right]-\mathrm{KL}\left[q(\boldsymbol{u}_*)||p(\boldsymbol{u}_*)\right], \quad \text{(C.8)}\end{aligned}$$

where $q_{\mathcal{C}}(\boldsymbol{u}_k)$ is the *contrastive* predictive GP posterior, whose derivation is provided in the next section C.1.2. Importantly, the ensemble bound in Eq. (C.8) is the one that we aim to maximise w.r.t. some variational parameters and hyperparameters. For a better comprehension of this point, we provide an extra-view of the bound and the presence of (fixed) local and (unfixed) global parameters in each term. See section C.1.3 for this.

C.1.1 Gaussian marginals for infinite-dimensional integral operators

The properties of Gaussian marginal distributions indicate that, having two *normal*-distributed random variables $\boldsymbol{a}$ and $\boldsymbol{b}$, its joint probability distribution is given by

$$p(\boldsymbol{a},\boldsymbol{b}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_a\\ \boldsymbol{\mu}_b\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab}\\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb}\end{bmatrix}\right),$$

and if we want to marginalize one of that variables out, such as $\int p(\boldsymbol{a},\boldsymbol{b})d\boldsymbol{b}$. It turns to be

$$\int p(\boldsymbol{a},\boldsymbol{b})d\boldsymbol{b} = p(\boldsymbol{a}) = \mathcal{N}(\boldsymbol{\mu}_a,\boldsymbol{\Sigma}_{aa}).$$

This same property is applicable to every derivation with GPs. In our case, it is the key point that we use to reduce the infinite-dimensional integral operators over the full stochastic processes. An example can be found in the expectation $\mathbb{E}_{p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*)}[\cdot]$ of Eq. (C.6). Its final

derivation to only integrate on $\boldsymbol{u}_k$ rather than on $\boldsymbol{f}_{\infty \neq \boldsymbol{u}_*}$ comes from

$$
\begin{aligned}
p(\boldsymbol{f}_{\infty \neq \boldsymbol{u}_*}|\boldsymbol{u}_*) &= p(\boldsymbol{f}_{\infty \neq \{\boldsymbol{u}_*, \boldsymbol{u}_k\}}, \boldsymbol{u}_k|\boldsymbol{u}_*) \\
&= \mathcal{N}\left(\begin{bmatrix} \boldsymbol{m}_{\boldsymbol{f}_{\infty \neq \{\boldsymbol{u}_*, \boldsymbol{u}_k\}}|\boldsymbol{u}_*} \\ \boldsymbol{m}_{\boldsymbol{u}_k|\boldsymbol{u}_*} \end{bmatrix}, \begin{bmatrix} \boldsymbol{Q}_{\boldsymbol{f}_{\infty \neq \{\boldsymbol{u}_*, \boldsymbol{u}_k\}}|\boldsymbol{u}_*} & \boldsymbol{Q}_{\boldsymbol{f}_{\infty \neq \{\boldsymbol{u}_*, \boldsymbol{u}_k\}}, \boldsymbol{u}_k|\boldsymbol{u}_*} \\ \boldsymbol{Q}_{\boldsymbol{u}_k, \boldsymbol{f}_{\infty \neq \{\boldsymbol{u}_*, \boldsymbol{u}_k\}}|\boldsymbol{u}_*} & \boldsymbol{Q}_{\boldsymbol{u}_k|\boldsymbol{u}_*} \end{bmatrix} \right),
\end{aligned}
$$

and if we marginalize over $\boldsymbol{f}_{\infty \neq \{\boldsymbol{u}_*, \boldsymbol{u}_k\}}|\boldsymbol{u}_*$, ends in the following reduction of the conditional prior expectation

$$
\begin{aligned}
\mathbb{E}_{p(\boldsymbol{f}_{\infty \neq \boldsymbol{u}_*}|\boldsymbol{u}_*)}\left[g(\boldsymbol{u}_k)\right] &= \int p(\boldsymbol{f}_{\infty \neq \boldsymbol{u}_*}|\boldsymbol{u}_*) g(\boldsymbol{u}_k) d\boldsymbol{f}_{\infty \neq \boldsymbol{u}_*} \\
&= \iint p(\boldsymbol{f}_{\infty \neq \{\boldsymbol{u}_*, \boldsymbol{u}_k\}}, \boldsymbol{u}_k|\boldsymbol{u}_*) g(\boldsymbol{u}_k) d\boldsymbol{f}_{\infty \neq \{\boldsymbol{u}_*, \boldsymbol{u}_k\}} d\boldsymbol{u}_k \\
&= \int p(\boldsymbol{u}_k|\boldsymbol{u}_*) g(\boldsymbol{u}_k) d\boldsymbol{u}_k = \mathbb{E}_{p(\boldsymbol{u}_k|\boldsymbol{u}_*)}\left[g(\boldsymbol{u}_k)\right], \quad \text{(C.9)}
\end{aligned}
$$

where we denote $g(\boldsymbol{u}_k) = \log\left(q_k(\boldsymbol{u}_k)/p_k(\boldsymbol{u}_k)\right)$ and we used

$$
\int p(\boldsymbol{f}_{\infty \neq \{\boldsymbol{u}_*, \boldsymbol{u}_k\}}, \boldsymbol{u}_k|\boldsymbol{u}_*) d\boldsymbol{f}_{\infty \neq \{\boldsymbol{u}_*, \boldsymbol{u}_k\}} = p(\boldsymbol{u}_k) = \mathcal{N}(\boldsymbol{m}_{\boldsymbol{u}_k|\boldsymbol{u}_*}, \boldsymbol{Q}_{\boldsymbol{u}_k|\boldsymbol{u}_*}).
$$

C.1.2 Contrastive posterior GP predictive

The *contrastive* predictive GP posterior distribution $q_{\mathcal{C}}(\boldsymbol{u}_k)$ is obtained from the *nested* integration in Eq. (C.8). We begin its derivation with the l.h.s. expectation term in Eq. (C.8), then

$$
\begin{aligned}
&\sum_{k=1}^{K} \mathbb{E}_{q(\boldsymbol{u}_*)}\left[\mathbb{E}_{p(\boldsymbol{u}_k|\boldsymbol{u}_*)}\left[\log\left(\frac{q_k(\boldsymbol{u}_k)}{p_k(\boldsymbol{u}_k)}\right)\right]\right] \\
&\qquad = \sum_{k=1}^{K} \iint q(\boldsymbol{u}_*) p(\boldsymbol{u}_k|\boldsymbol{u}_*) \log\left(\frac{q_k(\boldsymbol{u}_k)}{p_k(\boldsymbol{u}_k)}\right) d\boldsymbol{u}_k d\boldsymbol{u}_* \\
&\qquad = \sum_{k=1}^{K} \int \underbrace{\left(\int q(\boldsymbol{u}_*) p(\boldsymbol{u}_k|\boldsymbol{u}_*) d\boldsymbol{u}_*\right)}_{q_{\mathcal{C}}(\boldsymbol{u}_k)} \log\left(\frac{q_k(\boldsymbol{u}_k)}{p_k(\boldsymbol{u}_k)}\right) d\boldsymbol{u}_k, \quad \text{(C.10)}
\end{aligned}
$$

where the conditional GP prior distribution between the local inducing-inputs $\boldsymbol{u}_k$ and the global ones $\boldsymbol{u}_*$, is $p(\boldsymbol{u}_k|\boldsymbol{u}_*) = \mathcal{N}(\boldsymbol{u}_k|\boldsymbol{m}_{k|*}, \boldsymbol{Q}_{k|*})$ with

$$
\begin{aligned}
\boldsymbol{m}_{k|*} &= \mathbf{K}_{*k}^{\top} \mathbf{K}_{**}^{-1} \boldsymbol{u}_*, \\
\boldsymbol{Q}_{k|*} &= \mathbf{K}_k - \mathbf{K}_{*k}^{\top} \mathbf{K}_{**}^{-1} \mathbf{K}_{*k},
\end{aligned}
$$

and where covariance matrices are built from $[\mathbf{K}_{**}]_{m,n} := k(\boldsymbol{z}_m, \boldsymbol{z}_n)$ with $\boldsymbol{z}_m, \boldsymbol{z}_n \in \mathbb{R}^p$. Finally, the contrastive predictive GP posterior $q_{\mathcal{C}}(\boldsymbol{u}_k)$ can be computed from the expectation term in Eq. (C.10) as

$$
\int q(\boldsymbol{u}_*) p(\boldsymbol{u}_k|\boldsymbol{u}_*) d\boldsymbol{u}_* = q_{\mathcal{C}}(\boldsymbol{u}_k) = \mathcal{N}(\boldsymbol{u}_k|\boldsymbol{m}_{\mathcal{C}}, \boldsymbol{S}_{\mathcal{C}}), \quad \text{(C.11)}
$$

where the parameters $\boldsymbol{m}_{\mathcal{C}}$ and $\boldsymbol{S}_{\mathcal{C}}$ are

$$
\begin{aligned}
\boldsymbol{m}_{\mathcal{C}} &= \mathbf{K}_{*k}^{\top} \mathbf{K}_{**}^{-1} \boldsymbol{\mu}_*, \\
\boldsymbol{S}_{\mathcal{C}} &= \mathbf{K}_k - \mathbf{K}_{*k}^{\top} \mathbf{K}_{**}^{-1} (\boldsymbol{S}_* - \mathbf{K}_{**}) \mathbf{K}_{**}^{-1} \mathbf{K}_{*k}.
\end{aligned}
$$

C.1.3 Parameters in the lower ensemble bound

We approximate the global approximation to the GP posterior distribution as $q(f) \approx p(f|\mathcal{D})$. Additionally, we introduce the subset of global inducing-inputs $\boldsymbol{Z}_* = \{\boldsymbol{z}_m\}_{m=1}^M$ and their corresponding function evaluations are $\boldsymbol{u}_*$. Then, the *explicit* variational distribution given the pseudo-observations $\boldsymbol{u}_*$ is $q(\boldsymbol{u}_*) = \mathcal{N}(\boldsymbol{u}_*|\boldsymbol{\mu}_*, \boldsymbol{S}_*)$. Previously, we have obtained the list of objects $\mathcal{E} = \{\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_K\}$ without any specific order, where each $\mathcal{E}_k = \{\boldsymbol{\phi_k}, \boldsymbol{\psi_k}, \boldsymbol{Z}_k\}$, $\boldsymbol{\phi_k}$ being the corresponding local variational parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{S}_k$.

If we look to the ensemble lower bound in Eq. (C.8), we omitted the conditioning on both variational parameters and hyperparameters for clarity. However, to make this point clear, we will now rewritte Eq. (C.8) to show the influence of each parameter variable over each term in the global bound. We remark that $\{\boldsymbol{\phi_k}, \boldsymbol{\psi_k}\}_{k=1}^K$ are given and fixed, whilst $\{\boldsymbol{\phi}_*, \boldsymbol{\psi}_*\}$ are the variational parameters and hyperparameters that we aim to fit,

$$\mathcal{L}_{\mathcal{E}}(\boldsymbol{\phi}_*, \boldsymbol{\psi}_*) = \sum_{k=1}^{K} \mathbb{E}_{q_{\mathcal{C}}(\boldsymbol{u}_k|\boldsymbol{\phi}_*, \boldsymbol{\psi}_*)} \left[\log q_k(\boldsymbol{u}_k|\boldsymbol{\phi_k}) - \log p_k(\boldsymbol{u}_k|\boldsymbol{\psi_k})\right] - \mathrm{KL}\left[q(\boldsymbol{u}_*|\boldsymbol{\phi}_*)||p(\boldsymbol{u}_*|\boldsymbol{\psi}_*)\right].$$

We remind that the global variational parameters are $\boldsymbol{\phi}_* = \{\boldsymbol{\mu}_*, \boldsymbol{S}_*\}$, while the hyperparameters would correspond to $\boldsymbol{\psi}_* = \{\ell, \sigma_a\}$ in the case of using the vanilla *kernel*, with ℓ being the lengthscale and σ_a the amplitude variables. The notation of the local counterpart is equivalent.

The dependencies of parameters in our Pytorch implementation (https://github.com/pmorenoz/RecyclableGP) are clearly shown and evident from the code structure oriented to objects. It is also amenable for the introduction of new covariance functions and more structured variational approximations if needed.

C.2 Distributions and Expectations

To assure the future and easy reproducibility of our recyclable GP framework, we provide the exact expression of all distributions and expectations involved in the lower ensemble bound in Eq. (C.8).

C.2.1 Distributions

The log-distributions and distributions that appear in Eq. (C.8) are $\log q(\boldsymbol{u}_k)$, $\log p(\boldsymbol{u}_k)$, $q(\boldsymbol{u}_*)$, $p(\boldsymbol{u}_*)$ and $q_{\mathcal{C}}(\boldsymbol{u}_k)$. First, the computation of the logarithmic distributions is

$$\log q(\boldsymbol{u}_k) = \log\left(\mathcal{N}(\boldsymbol{u}_k|\boldsymbol{\mu}_k, \boldsymbol{S}_k)\right) = -\frac{1}{2}(\boldsymbol{u}_k - \boldsymbol{\mu}_k)^\top \boldsymbol{S}_k^{-1}(\boldsymbol{u}_k - \boldsymbol{\mu}_k) - \frac{1}{2}\log\det(2\pi\boldsymbol{S}_k),$$

$$\log p(\boldsymbol{u}_k) = \log\left(\mathcal{N}(\boldsymbol{u}_k|\boldsymbol{0}, \mathbf{K}_{kk})\right) = -\frac{1}{2}\boldsymbol{u}_k^\top \mathbf{K}_{kk}^{-1}\boldsymbol{u}_k - \frac{1}{2}\log\det(2\pi\mathbf{K}_{kk}),$$

while $q(\boldsymbol{u}_*)$ and $p(\boldsymbol{u}_*)$ are just $q(\boldsymbol{u}_*) = \mathcal{N}(\boldsymbol{u}_*|\boldsymbol{\mu}_*, \boldsymbol{S}_*)$ and $p(\boldsymbol{u}_*) = \mathcal{N}(\boldsymbol{u}_*|\boldsymbol{0}, \mathbf{K}_{**})$. The exact expression of the distribution $q_{\mathcal{C}}(\boldsymbol{u}_k)$ is provided in the section C.1.2.

C.2.2 Expectations

The K expectations in the l.h.s. term in Eq. (C.8) can be rewritten as

$$\begin{aligned}\sum_{k=1}^{K} \mathbb{E}_{q_{\mathcal{C}}(\boldsymbol{u}_k)} \left[\log q_k(\boldsymbol{u}_k) - \log p_k(\boldsymbol{u}_k)\right] \\ = \sum_{k=1}^{K} \left[\mathbb{E}_{q_{\mathcal{C}}(\boldsymbol{u}_k)} \left[\log q_k(\boldsymbol{u}_k)\right] - \mathbb{E}_{q_{\mathcal{C}}(\boldsymbol{u}_k)} \left[\log p_k(\boldsymbol{u}_k)\right]\right] \\ = \sum_{k=1}^{K} \left[\left\langle \log q_k(\boldsymbol{u}_k) \right\rangle_{q_{\mathcal{C}}(\boldsymbol{u}_k)} - \left\langle \log p_k(\boldsymbol{u}_k) \right\rangle_{q_{\mathcal{C}}(\boldsymbol{u}_k)}\right], \quad \text{(C.12)}\end{aligned}$$

where the k-th expectations over both $\log q_k(\boldsymbol{u}_k)$ and $\log p_k(\boldsymbol{u}_k)$ take the form

$$\left\langle \log q_k(\boldsymbol{u}_k) \right\rangle_{q_{\mathcal{C}}(\boldsymbol{u}_k)} = -\frac{1}{2}\left(\mathrm{Tr}\left(\boldsymbol{S}_k^{-1}\boldsymbol{S}_{\mathcal{C}}\right) + (\boldsymbol{m}_{\mathcal{C}} - \boldsymbol{\mu}_k)^{\top}\boldsymbol{S}_k^{-1}(\boldsymbol{m}_{\mathcal{C}} - \boldsymbol{\mu}_k) + \log\det\left(2\pi\boldsymbol{S}_k\right)\right),$$

$$\left\langle \log p_k(\boldsymbol{u}_k) \right\rangle_{q_{\mathcal{C}}(\boldsymbol{u}_k)} = -\frac{1}{2}\left(\mathrm{Tr}\left(\mathbf{K}_{kk}^{-1}\boldsymbol{S}_{\mathcal{C}}\right) + \boldsymbol{m}_{\mathcal{C}}^{\top}\mathbf{K}_{kk}^{-1}\boldsymbol{m}_{\mathcal{C}} + \log\det\left(2\pi\mathbf{K}_{kk}\right)\right).$$

C.3 Combined Ensemble Bounds with Unseen Data

As we already mentioned in the manuscript, there might be scenarios where it could be not necessary to distribute the whole dataset $\mathcal{D}$ in K local tasks or, for instance, a new *unseen* subset $k+1$ of observations might be available for processing. In such case, it is still possible to obtain a *combined* global solution that fits both to the local GP approximations and the new data. For clarity on this point, we rewrite the principal steps of the ensemble bound derivation in section A but without substituting all the log-likelihood terms by its Bayesian approximation, that is

$$\begin{aligned}\mathcal{L}_{\mathcal{E}} &= \mathbb{E}_{q(\boldsymbol{u}_*)}\left[\mathbb{E}_{p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*)}\left[\sum_{k=1}^{K}\log p(\boldsymbol{y}_k|f_\infty) + \log p(\boldsymbol{y}_{k+1}|f_\infty)\right] - \log\left(\frac{q(\boldsymbol{u}_*)}{p(\boldsymbol{u}_*)}\right)\right] \\ &= \mathbb{E}_{q(\boldsymbol{u}_*)}\left[\sum_{k=1}^{K}\mathbb{E}_{p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*)}\left[\log p(\boldsymbol{y}_k|f_\infty)\right] + \mathbb{E}_{p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*)}\left[\log p(\boldsymbol{y}_{k+1}|f_\infty)\right] - \log\left(\frac{q(\boldsymbol{u}_*)}{p(\boldsymbol{u}_*)}\right)\right] \\ &= \mathbb{E}_{q(\boldsymbol{u}_*)}\left[\sum_{k=1}^{K}\mathbb{E}_{p(\boldsymbol{u}_k|\boldsymbol{u}_*)}\left[\log\left(\frac{q_k(\boldsymbol{u}_k)}{p_k(\boldsymbol{u}_k)}\right)\right] + \mathbb{E}_{p(\boldsymbol{f}_{k+1}|\boldsymbol{u}_*)}\left[\log p(\boldsymbol{y}_{k+1}|\boldsymbol{f}_{k+1})\right] - \log\left(\frac{q(\boldsymbol{u}_*)}{p(\boldsymbol{u}_*)}\right)\right] \\ &= \sum_{k=1}^{K}\mathbb{E}_{q_{\mathcal{C}}(\boldsymbol{u}_k)}\left[\log q_k(\boldsymbol{u}_k) - \log p_k(\boldsymbol{u}_k)\right] + \sum_{i=1}^{N_{k+1}}\mathbb{E}_{q(\boldsymbol{f}_i)}\left[\log p(y_i|\boldsymbol{f}_i)\right] - \mathrm{KL}\left[q(\boldsymbol{u}_*)||p(\boldsymbol{u}_*)\right], \quad \text{(C.13)}\end{aligned}$$

where $q(\mathbf{f}_i)$ is the result of the integral $q(\mathbf{f}_i) = \int q(\boldsymbol{u}_*)p(\boldsymbol{f}_i|\boldsymbol{u}_*)d\boldsymbol{u}_*$ and we applied the factorisation to the *new* $(k+1)$-th expectation term as in Hensman et al. (2015a).

APPENDIX D

CONTINUAL MULTI-TASK GAUSSIAN PROCESSES

D.1 Complete derivation of continual lower bounds

To derive the *continual* lower bound for each iteration of the sequential process, we use the following expression

$$\begin{aligned}\log p(\boldsymbol{y}) &= \log \int p(\boldsymbol{y}|f_\infty)p(f_\infty)df_\infty = \log \int p(\boldsymbol{y}_{\text{new}}, \boldsymbol{y}_{\text{old}}|f_\infty)p(f_\infty)df_\infty \\ &= \log \int p(\boldsymbol{y}_{\text{new}}|f_\infty)p(\boldsymbol{y}_{\text{old}}|f_\infty)p(f_\infty)df_\infty \geq \mathcal{L}_{\mathcal{C}}, \end{aligned} \tag{D.1}$$

where we applied the Jensen's inequality. Then, the integral of the logarithmic distributions is

$$\begin{aligned}\mathcal{L}_{\mathcal{C}} &= \int \log p(\boldsymbol{y}_{\text{new}}|f_\infty)p(\boldsymbol{y}_{\text{old}}|f_\infty)p(f_\infty)df_\infty \\ &= \int q(f_\infty|\boldsymbol{\phi}_{\text{new}}) \log \frac{p(\boldsymbol{y}_{\text{new}}|f_\infty)p(\boldsymbol{y}_{\text{old}}|f_\infty)p(f_\infty)}{q(f_\infty|\boldsymbol{\phi}_{\text{new}})}df_\infty \\ &= \int q(f_\infty|\boldsymbol{\phi}_{\text{new}}) \log \frac{p(\boldsymbol{y}_{\text{new}}|f_\infty)q(f_\infty|\boldsymbol{\phi}_{\text{old}})p(f_\infty|\boldsymbol{\psi}_{\text{new}})}{p(f_\infty|\boldsymbol{\psi}_{\text{old}})q(f_\infty|\boldsymbol{\phi}_{\text{new}})}df_\infty \\ &= \int q(f_\infty|\boldsymbol{\phi}_{\text{new}}) \log \frac{p(\boldsymbol{y}_{\text{new}}|f_\infty)p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*, \boldsymbol{\psi}_{\text{old}})\widetilde{q}(\boldsymbol{u}_*|\boldsymbol{\phi}_{\text{old}})p(f_\infty|\boldsymbol{\psi}_{\text{new}})}{p(f_\infty|\boldsymbol{\psi}_{\text{old}})q(f_\infty|\boldsymbol{\phi}_{\text{new}})}df_\infty \\ &= \int q(f_\infty|\boldsymbol{\phi}_{\text{new}}) \log \frac{p(\boldsymbol{y}_{\text{new}}|f_\infty)p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*, \boldsymbol{\psi}_{\text{old}})\widetilde{q}(\boldsymbol{u}_*|\boldsymbol{\phi}_{\text{old}})p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*, \boldsymbol{\psi}_{\text{new}})p(\boldsymbol{u}_*|\boldsymbol{\psi}_{\text{new}})}{p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*, \boldsymbol{\psi}_{\text{old}})p(\boldsymbol{u}_*|\boldsymbol{\psi}_{\text{old}})p(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}|\boldsymbol{u}_*, \boldsymbol{\psi}_{\text{new}})q(\boldsymbol{u}_*|\boldsymbol{\phi}_{\text{new}})}df_\infty \\ &= \int q(f_\infty|\boldsymbol{\phi}_{\text{new}}) \log \frac{p(\boldsymbol{y}_{\text{new}}|f_\infty)\widetilde{q}(\boldsymbol{u}_*|\boldsymbol{\phi}_{\text{old}})p(\boldsymbol{u}_*|\boldsymbol{\psi}_{\text{new}})}{p(\boldsymbol{u}_*|\boldsymbol{\psi}_{\text{old}})q(\boldsymbol{u}_*|\boldsymbol{\phi}_{\text{new}})}df_\infty \\ &= \int q(f_\infty|\boldsymbol{\phi}_{\text{new}}) \log p(\boldsymbol{y}_{\text{new}}|f_\infty)df_\infty - \int q(f_\infty|\boldsymbol{\phi}_{\text{new}}) \log \frac{q(\boldsymbol{u}_*|\boldsymbol{\phi}_{\text{new}})}{p(\boldsymbol{u}_*|\boldsymbol{\psi}_{\text{new}})}df_\infty \\ &+ \int q(f_\infty|\boldsymbol{\phi}_{\text{new}}) \log \frac{\widetilde{q}(\boldsymbol{u}_*|\boldsymbol{\phi}_{\text{old}})}{p(\boldsymbol{u}_*|\boldsymbol{\psi}_{\text{old}})}df_\infty \\ &= \int q(\boldsymbol{f}_{\infty\neq\{\boldsymbol{f}_{\text{new}},\boldsymbol{u}_*\}}, \boldsymbol{f}_{\text{new}}, \boldsymbol{u}_*|\boldsymbol{\phi}_{\text{new}}) \log p(\boldsymbol{y}_{\text{new}}|\boldsymbol{f}_{\text{new}})\boldsymbol{f}_{\infty\neq\{\boldsymbol{f}_{\text{new}},\boldsymbol{u}_*\}}d\boldsymbol{f}_{\text{new}}d\boldsymbol{u}_* \\ &- \int q(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}, \boldsymbol{u}_*|\boldsymbol{\phi}_{\text{new}}) \log \frac{q(\boldsymbol{u}_*|\boldsymbol{\phi}_{\text{new}})}{p(\boldsymbol{u}_*|\boldsymbol{\psi}_{\text{new}})}d\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}d\boldsymbol{u}_* \\ &+ \int q(\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}, \boldsymbol{u}_*|\boldsymbol{\phi}_{\text{new}}) \log \frac{\widetilde{q}(\boldsymbol{u}_*|\boldsymbol{\phi}_{\text{old}})}{p(\boldsymbol{u}_*|\boldsymbol{\psi}_{\text{old}})}d\boldsymbol{f}_{\infty\neq\boldsymbol{u}_*}d\boldsymbol{u}_* \\ &= \int q(\boldsymbol{u}_*|\boldsymbol{\phi}_{\text{new}})p(\boldsymbol{f}_{\text{new}}|\boldsymbol{u}_*) \log p(\boldsymbol{y}_{\text{new}}|\boldsymbol{f}_{\text{new}})d\boldsymbol{f}_{\text{new}}d\boldsymbol{u}_* - \int q(\boldsymbol{u}_*|\boldsymbol{\phi}_{\text{new}}) \log \frac{q(\boldsymbol{u}_*|\boldsymbol{\phi}_{\text{new}})}{p(\boldsymbol{u}_*|\boldsymbol{\psi}_{\text{new}})}d\boldsymbol{u}_* \\ &+ \int q(\boldsymbol{u}_*|\boldsymbol{\phi}_{\text{new}}) \log \frac{q(\boldsymbol{u}_*|\boldsymbol{\phi}_{\text{new}})\widetilde{q}(\boldsymbol{u}_*|\boldsymbol{\phi}_{\text{old}})}{q(\boldsymbol{u}_*|\boldsymbol{\phi}_{\text{new}})p(\boldsymbol{u}_*|\boldsymbol{\psi}_{\text{old}})}d\boldsymbol{u}_*, \end{aligned} \tag{D.2}$$

where we assumed $\boldsymbol{u}_*$ to be the new subset of inducing-points $\boldsymbol{u}_{\text{new}}$, then

$$
\begin{aligned}
&= \int q(\boldsymbol{f}_{\text{new}}) \log p(\boldsymbol{y}_{\text{new}}|\boldsymbol{f}_{\text{new}}) d\boldsymbol{f}_{\text{new}} - \int q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}}) \log \frac{q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})}{p(\boldsymbol{u}_{\text{new}}|\boldsymbol{\psi}_{\text{new}})} d\boldsymbol{u}_{\text{new}} \\
&+ \int q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}}) \log \frac{q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})}{p(\boldsymbol{u}_{\text{new}}|\boldsymbol{\psi}_{\text{old}})} d\boldsymbol{u}_{\text{new}} - \int q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}}) \log \frac{q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})}{\widetilde{q}(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{old}})} d\boldsymbol{u}_{\text{new}} \\
&= \mathbb{E}_{q(\boldsymbol{f}_{\text{new}})}[\log p(\boldsymbol{y}_{\text{new}}|\boldsymbol{f}_{\text{new}})] - \text{KL}[q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})||p(\boldsymbol{u}_{\text{new}}|\boldsymbol{\psi}_{\text{new}})] \\
&+ \text{KL}[q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})||p(\boldsymbol{u}_{\text{new}}|\boldsymbol{\psi}_{\text{old}})] - \text{KL}[q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})||\widetilde{q}(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}})]. \qquad \text{(D.3)}
\end{aligned}
$$

It is also important to rely on the variational expectation terms for the likelihood distributions where $q(\boldsymbol{f}_{\text{new}})$ intervenes. Particularly, we can take the explicit vector values $\boldsymbol{u}_{\text{new}}$ for the implicit inducing points notation $\boldsymbol{u}_*$. Then, the general expectation integral takes the form

$$
\begin{aligned}
\mathbb{E}_{q(\boldsymbol{f}_{\text{new}})}[\log p(\boldsymbol{y}_{\text{new}}|\boldsymbol{f}_{\text{new}})] &= \int q(\boldsymbol{u}_*|\boldsymbol{\phi}_{\text{new}}) p(\boldsymbol{f}_{\text{new}}|\boldsymbol{u}_*) \log p(\boldsymbol{y}_{\text{new}}|\boldsymbol{f}_{\text{new}}) d\boldsymbol{f}_{\text{new}} d\boldsymbol{u}_* \\
&= \int q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}}) p(\boldsymbol{f}_{\text{new}}|\boldsymbol{u}_{\text{new}}) \log p(\boldsymbol{y}_{\text{new}}|\boldsymbol{f}_{\text{new}}) d\boldsymbol{f}_{\text{new}} d\boldsymbol{u}_{\text{new}} \\
&= \int q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}}) p(\boldsymbol{f}_{\text{new}}|\boldsymbol{u}_{\text{new}}) d\boldsymbol{u}_{\text{new}} \log p(\boldsymbol{y}_{\text{new}}|\boldsymbol{f}_{\text{new}}) d\boldsymbol{f}_{\text{new}} \\
&= \int q(\boldsymbol{f}_{\text{new}}) \log p(\boldsymbol{y}_{\text{new}}|\boldsymbol{f}_{\text{new}}) d\boldsymbol{f}_{\text{new}}, \qquad \text{(D.4)}
\end{aligned}
$$

and considering we denote $q(\boldsymbol{f}_{\text{new}})$ as the expected variational distribution over the output vector $\mathbf{f}_{\text{new}}$, that can be analytically calculated as follows

$$
\begin{aligned}
q(\boldsymbol{f}_{\text{new}}) &= \int q(\boldsymbol{u}_{\text{new}}|\boldsymbol{\phi}_{\text{new}}) p(\boldsymbol{f}_{\text{new}}|\boldsymbol{u}_{\text{new}}) d\boldsymbol{u}_{\text{new}} \\
&= \mathcal{N}(\boldsymbol{f}_{\text{new}}|\mathbf{K}_{\boldsymbol{f}_{\text{new}}\boldsymbol{u}_{\text{new}}} \mathbf{K}_{\boldsymbol{u}_{\text{new}}\boldsymbol{u}_{\text{new}}}^{-1} \boldsymbol{\mu}_{\text{new}}, \mathbf{K}_{\boldsymbol{f}_{\text{new}}\boldsymbol{f}_{\text{new}}} \\
&+ \mathbf{K}_{\boldsymbol{f}_{\text{new}}\boldsymbol{u}_{\text{new}}} \mathbf{K}_{\boldsymbol{u}_{\text{new}}\boldsymbol{u}_{\text{new}}}^{-1} (\mathbf{S}_{\text{new}} - \mathbf{K}_{\boldsymbol{u}_{\text{new}}\boldsymbol{u}_{\text{new}}}) \mathbf{K}_{\boldsymbol{u}_{\text{new}}\boldsymbol{u}_{\text{new}}}^{-1} \mathbf{K}_{\boldsymbol{f}_{\text{new}}\boldsymbol{u}_{\text{new}}}^{\top}). \qquad \text{(D.5)}
\end{aligned}
$$

D.2 Dimensionality reduction of $p(f_\infty)$ via Gaussian marginals.

We use the properties of Gaussian marginals to reduce infinite dimensional distributions $p(f_\infty)$ in the continual approach. This process is applied for both GP priors $p(f_\infty)$, one w.r.t. hyperparameters $\boldsymbol{\psi}_{\text{new}}$ and other w.r.t. $\boldsymbol{\psi}_{\text{old}}$, and the Gaussian variational distribution $q(f_\infty)$. We also assume that if the generative process of the latent functions is $f \sim p(f_\infty)$, then it also holds

$$
\begin{bmatrix} \boldsymbol{f}_{\infty \neq \boldsymbol{u}_{\text{new}}} \\ \boldsymbol{u}_{\text{new}} \end{bmatrix} \sim p(\boldsymbol{f}_{\infty \neq \boldsymbol{u}_{\text{new}}}, \boldsymbol{u}_{\text{new}}),
$$

where the multivariate Gaussian distribution $p(f_\infty) = p(\boldsymbol{f}_{\infty \neq \boldsymbol{u}_{\text{new}}}, \boldsymbol{u}_{\text{new}})$ has the following $\mathbf{K}$ and $\boldsymbol{\mu}$ parameters

$$
p(\boldsymbol{f}_{\neq \boldsymbol{u}_{\text{new}}}, \boldsymbol{u}_{\text{new}}) = \mathcal{N}\Big(\begin{bmatrix} \boldsymbol{\mu}_{\boldsymbol{f} \infty \neq \inf \boldsymbol{u}_{\text{new}}} \\ \boldsymbol{\mu}_{\boldsymbol{u}_{\text{new}}} \end{bmatrix}, \begin{bmatrix} \mathbf{K}_{\boldsymbol{f}_{\infty \neq \mathbf{u}_{\text{new}}} f_{\infty \neq \mathbf{u}_{\text{new}}}} & \mathbf{K}_{\boldsymbol{f}_{\infty \neq \mathbf{u}_{\text{new}}} \boldsymbol{u}_{\text{new}}} \\ \mathbf{K}_{\boldsymbol{u}_{\text{new}} \boldsymbol{f}_{\infty \neq \mathbf{u}_{\text{new}}}} & \mathbf{K}_{\boldsymbol{u}_{\text{new}} \boldsymbol{u}_{\text{new}}} \end{bmatrix} \Big),
$$

and we therefore, may apply the marginalization $p(\boldsymbol{u}_{\text{new}})$ to obtain the target Gaussian distribution

$$
\int p(\boldsymbol{f}_{\infty \neq \boldsymbol{u}_{\text{new}}}, \boldsymbol{u}_{\text{new}}) d\boldsymbol{f}_{\infty \neq \boldsymbol{u}_{\text{new}}} = p(\boldsymbol{u}_{\text{new}}) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{u}_{\text{new}}}, \mathbf{K}_{\boldsymbol{u}_{\text{new}} \boldsymbol{u}_{\text{new}}}).
$$

APPENDIX E

ADDITIONAL EXPERIMENTS

E.1 Continual Multi-output Synchronous Channels

In this additional multi-output experiment with toy data, we are interested into jointly performing multi-task non-linear regression over two output Gaussian channels with different likelihood noise parameters. The underlying linear mixing of the latent functions is assumed to follow a LMC structure that we also aim to infer it in an online manner. The number of true latent functions is $Q = 2$ and we generate them using a linear combination of sinusoidal signals. In this case, we have artificially split the dataset into five batches of non-overlapping samples that are delivered sequentially at the same time-step on both channels. In Fig. E.1, we show three captures of the learning process for this experiment.

E.2 Continual Multi-output Asynchronous Channels

This experiment is of particular importance for the demonstration of the multi-output model performance under asymmetric incoming channels. Particularly, we consider the same dataset as in the synchronous scenario but introducing an asymmetric observation process over the incoming channels data by the learning system. That is, at each time-step, only one of the two channels delivers output-input samples. In the next step, the observation channel switches and new incoming data appears on the other one. This observation procedure is depicted in Fig. E.2.

The continual inference process is possible due to the latent functions $\mathcal{U}$ lie in a different layer than the output observations. Hence, the inducing points can be positioned across the input domain within the emergence of new samples in any of the output channels. The number of initial inducing points is $M_q = 4$ per channel, and double per time-step iteration.

E.3 Multi-channel sensors for Human Motion

For the last experiment of this thesis on multi-output regression with real-world data, we consider the MOCAP dataset.[1] The data consists of raw multi-channel traces from sensors monitoring human motion. In particular, we select the first individual (id. number 01) in the *walking* activity example. We aim to exploit the benefits of multi-task GPs rather that using a single-output GP per sensor. It is demonstrated that by exploiting such correlations between channels, multiple-output data are better modelled (Bonilla et al., 2008). From all available sensors in the human body, we consider three of them whose oscillation phase does not coincide: the *left wrist*, the *right wrist* and at the *right femur*. Each channel provides a number of $N = 343$ samples corresponding to the vertical axis values recorded by the sensors. For the experiment, we setup an initial amount of $M = 10$ inducing inputs in order to obtain a reliable precision. We increase the M twice per recursive iteration. Moreover, the number

[1] MOCAP datasets are available at http://mocap.cs.cmu.edu/.

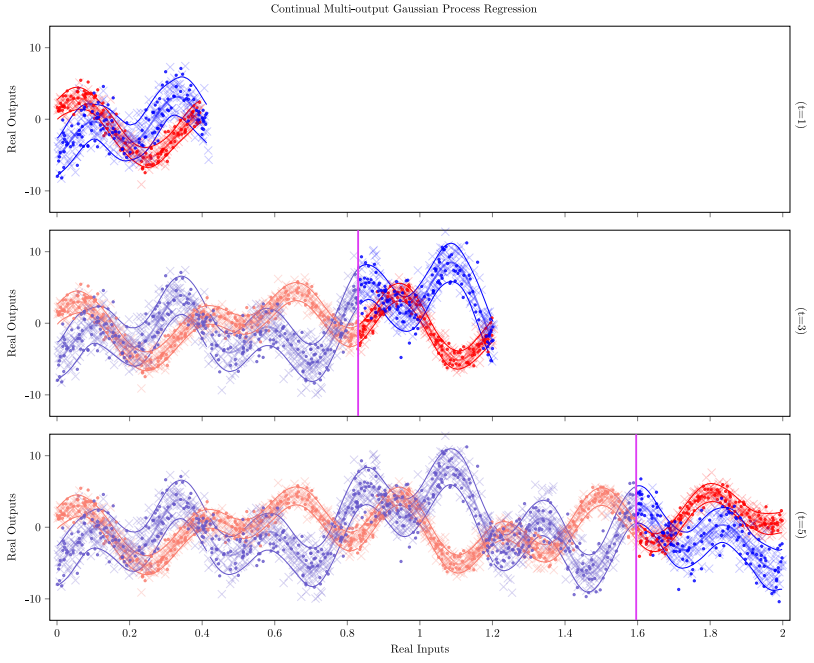


Figure E.1: Results for temporal modeling of multi-output real-valued data. Two channels are jointly model using the continual learning approach aforementioned for multi-output GP regression. The pink line indicates the limiting point between the novel observed samples and the past data that we avoid to revisit. All inducing inputs are positioned over the Q underlying latent functions that are later combined to obtain the output parameter functions. Both channels are trained together in a synchronous manner. The Q subsets of inducing-inputs are not plotted for a reason of clarity.

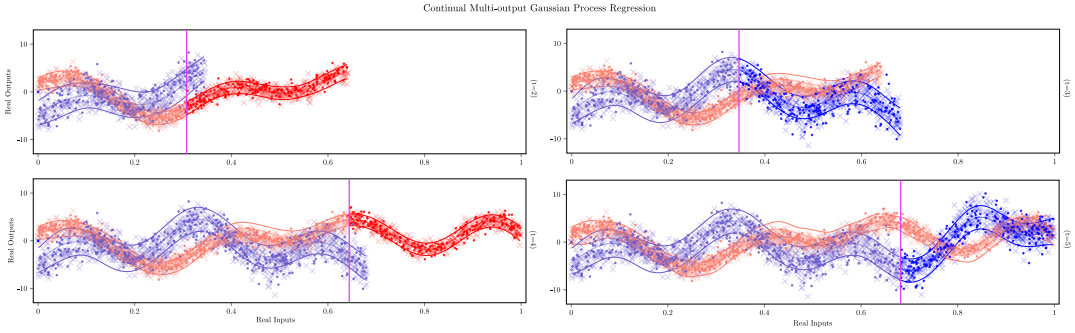


Figure E.2: In contrast to Fig. E.1, we apply the continual GP approach to model multi-channel sequential data that is observed in an asynchronous manner, that is, samples might appear at different time steps from different outputs in unobserved input regions. From left to right and from top to down, we represent the learning process at four consecutive time-steps ($t = 2$, $t = 3$, $t = 4$ and $t = 5$). Past data is plotted using grey scaled colors.

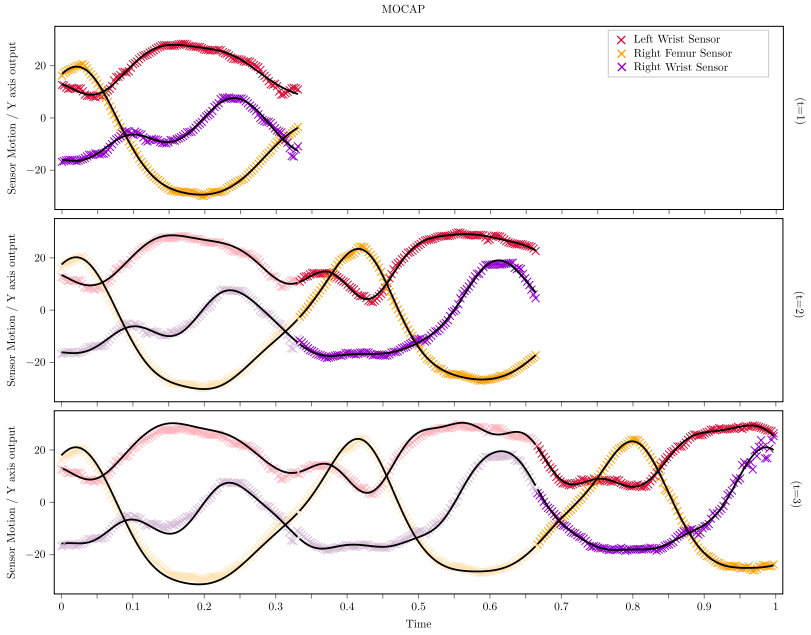


Figure E.3: MOCAP dataset. Multi-output GP regression over three sequential channels. Each channel corresponds to the Y axis output values of a sensor in a walking motion capture experiment. Black curves correspond to the mean of the posterior predictive distribution at each time-step for the whole input space. Gray scaled colors correspond to non-revisited data samples.

of latent functions in the multi-output GP prior is $Q = 3$. Both latent function values and the underlying linear mixing coefficients are initialized at random at each time-step.

BIBLIOGRAPHY

R. P. Adams and D. J. C. MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

D. Agudelo-España, S. Gomez-Gonzalez, S. Bauer, B. Schölkopf, and J. Peters. Bayesian online prediction of change points. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 320–329. PMLR, 2020.

T. Aledavood, S. Lehmann, and J. Saramäki. Digital daily cycles of individuals. *Frontiers in Physics*, 3(October):1–7, 2015a.

T. Aledavood, E. López, S. G. B. Roberts, F. Reed-Tsochas, E. Moro, R. I. M. Dunbar, and J. Saramäki. Daily rhythms in mobile telephone communication. *PLoS ONE*, 10(9):1–14, 2015b.

M. Alvarez and N. D. Lawrence. Sparse convolved Gaussian processes for multi-output regression. In *Advances in Neural Information Processing Systems (NIPS)*, pages 57–64, 2009.

M. Álvarez, D. Luengo, M. Titsias, and N. Lawrence. Variational inducing kernels for sparse convolved multiple output Gaussian processes. *arXiv preprint arXiv:0912.3268*, 2009.

M. Álvarez, D. Luengo, M. Titsias, and N. Lawrence. Efficient multioutput Gaussian processes through variational inducing kernels. In *Artificial Intelligence and Statistics (AISTATS)*, pages 25–32, 2010.

M. A. Álvarez and N. D. Lawrence. Computationally efficient convolved multiple output Gaussian processes. *Journal of Machine Learning Research (JMLR)*, 12(May):1459–1500, 2011.

M. A. Alvarez, L. Rosasco, N. D. Lawrence, et al. Kernels for vector-valued functions: A review. *Foundations and Trends in Machine Learning*, 4(3):195–266, 2012.

C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pages 1152–1174, 1974.

Y. Barak-Corren, V. M. Castro, S. Javitt, A. G. Hoffnagle, Y. Dai, R. H. Perlis, M. K. Nock, J. W. Smoller, and B. Y. Reis. Predicting suicidal behavior from longitudinal electronic health records. *American Journal of Psychiatry*, 174(2):154–162, 2017.

I. Barnett and J.-P. Onnela. Inferring mobility measures from gps traces with missing data. *Biostatistics*, 21(2):e98–e112, 2020.

I. Barnett, J. Torous, P. Staples, L. Sandoval, M. Keshavan, and J.-P. Onnela. Relapse prediction in schizophrenia through digital phenotyping: A pilot study. *Neuropsychopharmacology*, 2018.

M. L. Barrigón, S. Berrouiguet, J. J. Carballo, C. Bonal-Giménez, P. Fernández-Navarro, B. Pfang, D. Delgado-Gómez, P. Courtet, F. Aroca, J. Lopez-Castroman, et al. User profiles of an electronic mental health tool for ecological momentary assessment: MEmind. *International Journal of Methods in Psychiatric Research*, 26(1):e1554, 2017.

J. Bayer, C. Osendorfer, S. Diot-Girard, T. Rückstiess, and S. Urban. Climin: A pythonic framework for gradient-based function optimization. *TUM Tech. Report*, 2016.

M. J. Beal. Variational algorithms for approximate Bayesian inference. *Ph. D. Thesis, University College London*, 2003.

J. B. Begole, J. C. Tang, and R. Hill. Rhythm modeling, visualizations and applications. pages 11–20, 2003.

R. E. Bellman. *Adaptive control processes: A guided tour*. Princeton University Press, 1961.

S. Berrouiguet, M. Gravey, M. Le Galudec, Z. Alavi, and M. Walter. Post-acute crisis text messaging outreach for suicide prevention: A pilot study. *Psychiatry Research*, 217(3):154–157, 2014.

S. Berrouiguet, D. Ramírez, M. L. Barrigón, P. Moreno-Muñoz, R. Carmona, E. Baca-García, and A. Artés-Rodríguez. Combining Continuous Smartphone Native Sensors Data Capture and Unsupervised Data Mining Techniques to Detect Changes in Behavior: A Case Series of the Evidence-Based Behavior (eB2) Study. *JMIR MHealth and UHealth*, 2018.

S. Berrouiguet, M. L. Barrigón, J. L. Castroman, P. Courtet, A. Artés-Rodríguez, and E. Baca-García. Combining mobile-health (mHealth) and artificial intelligence (AI) methods to avoid suicide attempts: The smartcrises study protocol. *BMC Psychiatry*, 19(1):1–9, 2019.

C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

N. Bolger, A. Davis, and E. Rafaeli. Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, 54(1):579–616, 2003.

E. V. Bonilla, K. M. Chai, and C. Williams. Multi-task Gaussian process prediction. In *Advances in Neural Information Processing Systems (NIPS)*, pages 153–160, 2008.

R. F. Bordley. A multiplicative formula for aggregating probability assessments. *Management Science*, 28(10):1137–1148, 1982.

P. Boyle and M. Frean. Dependent Gaussian processes. *Advances in Neural Information Processing Systems (NeurIPS)*, 17:217–224, 2004.

J. V. Braun, R. Braun, and H.-G. Müller. Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika*, 87(2):301–314, 2000.

T. D. Bui and R. E. Turner. Tree-structured Gaussian process approximations. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2213–2221, 2014.

T. D. Bui, C. V. Nguyen, and R. E. Turner. Streaming sparse Gaussian process approximations. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3299–3307, 2017a.

T. D. Bui, J. Yan, and R. E. Turner. A unifying framework for Gaussian process pseudo-point approximations using power expectation propagation. *Journal of Machine Learning Research (JMLR)*, 18(1):3649–3720, 2017b.

T. D. Bui, C. V. Nguyen, S. Swaroop, and R. E. Turner. Partitioned variational inference: A unified framework encompassing federated and continual learning. *arXiv preprint arXiv:1811.11206*, 2018.

D. R. Burt, C. E. Rasmussen, and M. Van der Wilk. Rates of convergence for sparse variational Gaussian process regression. In *International Conference on Machine Learning (ICML)*, pages 862–871, 2019.

D. R. Burt, C. E. Rasmussen, and M. van der Wilk. Convergence of sparse variational inference in Gaussian processes regression. *Journal of Machine Learning Research*, 21(131):1–63, 2020.

F. Calabrese, D. Dahlem, A. Gerber, D. Paul, X. Chen, J. Rowland, C. Rath, and C. Ratti. The connected states of America: Quantifying social radii of influence. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 223–230. IEEE, 2011.

L. Canzian and M. Musolesi. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. *International Joint Conference on Pervasive and Ubiquitous Computing (UBICOMP)*, 2015.

Y. Cao and D. J. Fleet. Generalized product of experts for automatic and principled fusion of Gaussian process predictions. *arXiv preprint arXiv:1410.7827*, 2014.

O. Cappé and E. Moulines. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613, 2009.

A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Systems with Applications*, 39 (12):10873–10888, 2012.

K. M. A. Chai. Variational multinomial logit Gaussian process. *Journal of Machine Learning Research (JMLR)*, 13:1745–1808, 2012.

J. Chen and A. K. Gupta. Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical Association*, 92(438):739–747, 1997.

C.-A. Cheng and B. Boots. Incremental variational sparse Gaussian process regression. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4410–4418, 2016.

L.-F. Cheng, G. Darnell, C. Chivers, M. E. Draugelis, K. Li, and B. E. Engelhardt. Sparse multi-output Gaussian processes for medical time series prediction. *arXiv preprint arXiv:1703.09112*, 2017.

T. Choudhury, G. Borriello, S. Consolvo, D. Haehnel, B. Harrison, B. Hemingway, J. Hightower, P. Pedja, K. Koscher, A. LaMarca, et al. The mobile sensing platform: An embedded activity recognition system. *IEEE Pervasive Computing*, 7(2):32–41, 2008.

L. Csató and M. Opper. Sparse on-line Gaussian processes. *Neural Computation*, 14(3):641–668, 2002.

Z. Dai, M. A. Álvarez, and N. Lawrence. Efficient modeling of latent information in supervised learning using Gaussian processes. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5131–5139, 2017.

A. Damianou and N. D. Lawrence. Semi-described and semi-supervised learning with Gaussian processes. In *Uncertainty in Artificial Intelligence (UAI)*, pages 228–237, 2015.

M. Deisenroth and J. W. Ng. Distributed Gaussian processes. In *International Conference on Machine Learning (ICML)*, pages 1481–1490, 2015.

A. Dezfouli and E. V. Bonilla. Scalable inference for Gaussian process models with black-box likelihoods. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1414–1422, 2015.

N. Durrande, J. Hensman, M. Rattray, and N. D. Lawrence. Detecting periodicities with Gaussian processes. *PeerJ Computer Science*, 2:e50, 2016.

N. Eagle and A. S. Pentland. Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4):255–268, 2006.

N. Eagle and A. S. Pentland. Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7):1057–1066, 2009.

P. Fearnhead. Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16(2):203–213, 2006.

P. Fearnhead and P. Clifford. On-line inference for hidden Markov models via particle filters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(4):887–899, 2003.

P. Fearnhead and Z. Liu. On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):589–605, 2007.

J. Firth, J. Torous, and A. R. Yung. Ecological momentary assessment and beyond: the rising interest in e-mental health research. *Journal of Psychiatric Research*, 80:3–4, 2016.

V. Fortuin, G. Rätsch, and S. Mandt. Multivariate time series imputation with variational autoencoders. *arXiv preprint arXiv:1907.04155*, 2019.

V. Fortuin, D. Baranchuk, G. Rätsch, and S. Mandt. Gp-vae: Deep probabilistic time series imputation. In *Artificial Intelligence and Statistics (AISTATS)*, pages 1651–1661, 2020.

Y. Gal, M. Van Der Wilk, and C. E. Rasmussen. Distributed variational inference in sparse Gaussian process regression and latent variable models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3257–3265, 2014.

M. Garbarino, M. Lai, D. Bender, R. W. Picard, and S. Tognetti. Empatica E3 – A wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *2014 4th International Conference on Wireless Mobile Communication and Healthcare-Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH)*, pages 39–42. IEEE, 2014.

Z. Ghahramani and M. I. Jordan. Supervised learning from incomplete data via an EM approach. In *Advances in Neural Information Processing Systems (NIPS)*, 1994.

A. Girard, C. E. Rasmussen, J. Quiñonero Candela, and R. Murray-Smith. Gaussian process priors with uncertain inputs - Application to multiple-step ahead time series forecasting. In *Advances in Neural Information Processing Systems (NIPS)*, pages 545–552, 2003.

M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

T. L. Griffiths and Z. Ghahramani. The Indian buffet process: An introduction and review. *Journal of Machine Learning Research (JMLR)*, 12(4), 2011.

J. D. Hadfield et al. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software*, 33(2):1–22, 2010.

Z. Harchaoui, E. Moulines, and F. R. Bach. Kernel change-point analysis. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

J. Harrison, A. Sharma, C. Finn, and M. Pavone. Continuous meta-learning without tasks. *arXiv preprint arXiv:1912.08866*, 2019.

M. Heinonen, H. Mannerström, J. Rousu, S. Kaski, and H. Lähdesmäki. Non-stationary Gaussian process regression with hamiltonian Monte Carlo. In *Artificial Intelligence and Statistics*, 2016.

R. Henao and O. Winther. PASS-GP: Predictive active set selection for Gaussian processes. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 148–153, 2010.

J. Hensman, M. Rattray, and N. D. Lawrence. Fast variational inference in the conjugate exponential family. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2888–2896, 2012.

J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In *Uncertainty in Artificial Intelligence (UAI)*, pages 282–290, 2013a.

J. Hensman, N. D. Lawrence, and M. Rattray. Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. *BMC Bioinformatics*, 14(1):252, 2013b.

J. Hensman, A. G. d. G. Matthews, and Z. Ghahramani. Scalable variational Gaussian process classification. In *Artificial Intelligence and Statistics (AISTATS)*, pages 351–360, 2015a.

J. Hensman, M. Rattray, and N. D. Lawrence. Fast nonparametric clustering of structured time-series. *IEEE TPAMI*, 37(2):383–393, 2015b.

D. M. Higdon. *Space and space-time modelling using process convolutions*. 2002.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research (JMLR)*, 14(1):1303–1347, 2013.

M. Höhle. Online change-point detection in categorical time series. *Statistical Modelling and Regression Structures*, pages 377–397, 2010.

M. Husky, E. Olié, S. Guillaume, C. Genty, J. Swendsen, and P. Courtet. Feasibility and validity of ecological momentary assessment in the investigation of suicide risk. *Psychiatry Research*, 220 (1-2):564–570, 2014.

B. Jackson, J. D. Sargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, E. Gwin, P. Sangtrakulcharoen, L. Tan, and T. T. Tsai. An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, (12):105–108, 2005.

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

A. Jalal, S. Kamal, and D. Kim. Individual detection-tracking-recognition using depth activity images. In *2015 12th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pages 450–455. IEEE, 2015.

S. L. James, D. Abate, K. H. Abate, S. M. Abay, C. Abbafati, N. Abbasi, H. Abbastabar, F. Abd-Allah, J. Abdela, A. Abdelalim, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: A systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159): 1789–1858, 2018.

O. R. JK and F. WJ. *Numerical Bayesian methods applied to signal processing*. Springer, 1996.

J. Joseph, F. Doshi-Velez, A. S. Huang, and N. Roy. A Bayesian nonparametric approach to modeling motion patterns. *Autonomous Robots*, 31(4), 2011.

A. G. Journel and C. J. Huijbregts. *Mining Geostatistics*. Academic Press, 1978.

R. C. Kessler, P. Berglund, G. Borges, M. Nock, and P. S. Wang. Trends in suicide ideation, plans, gestures, and attempts in the United States, 1990-1992 to 2001-2003. *Jama*, 293(20):2487–2495, 2005.

M. E. E. Khan, G. Bouchard, K. P. Murphy, and B. M. Marlin. Variational bounds for mixed-data factor analysis. *Advances in Neural Information Processing Systems (NIPS)*, 23:1108–1116, 2010.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, 1982.

A. Klami, S. Virtanen, and S. Kaski. Bayesian exponential family projections for coupled data sources. *arXiv preprint arXiv:1203.3489*, 2012.

J. Knoblauch and T. Damoulas. Spatio-temporal Bayesian on-line changepoint detection with model selection. In *International Conference on Machine Learning (ICML)*, pages 2718–2727. PMLR, 2018.

K. Kroenke, R. L. Spitzer, and J. B. Williams. The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613, 2001.

L. I. Kuncheva. Change detection in streaming multivariate data using likelihood detectors. *IEEE Transactions on Knowledge and Data Engineering*, 25(5):1175–1180, 2011.

M. E. Larsen, T. W. Boonstra, P. J. Batterham, B. O'Dea, C. Paris, and H. Christensen. We feel: Mapping emotion on Twitter. *IEEE Journal of Biomedical and Health Informatics*, 19(4): 1246–1252, July 2015.

M. Lázaro-Gredilla and M. Titsias. Variational heteroscedastic Gaussian process regression. In *International Conference on Machine Learning (ICML)*, pages 841–848, 2011.

S. Li, Y. Xie, H. Dai, and L. Song. M-statistic for kernel change-point detection. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

A. Lima, R. Stanojevic, D. Papagiannaki, P. Rodriguez, and M. C. González. Understanding individual routing behavior. *Journal of the Royal Society Interface*, 13, 2016.

C. Ma, S. Tschiatschek, R. Turner, J. M. Hernández-Lobato, and C. Zhang. VAEM: A deep generative model for heterogeneous mixed type data. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020.

M. Ma, H. Fan, and K. M. Kitani. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1894–1903, 2016.

A. Madan, M. Cebrian, D. Lazer, and A. Pentland. Social sensing for epidemiological behavior change. pages 291–300, 2010.

L. Martino and V. Elvira. Metropolis sampling. *Wiley StatsRef: Statistics Reference Online*, 2017.

L. Martino, J. Read, and D. Luengo. Independent doubly adaptive rejection Metropolis sampling within Gibbs sampling. *IEEE Transactions on Signal Processing*, 63(12):3123–3138, 2015.

L. Martino, V. Elvira, and G. Camps-Valls. The recycling Gibbs sampler for efficient learning. *Digital Signal Processing*, 74:1–13, 2018.

L. Marzano, A. Bardill, B. Fields, K. Herd, D. Veale, N. Grey, and P. Moran. The application of mHealth to mental health: Opportunities and challenges. *The Lancet Psychiatry*, 2(10):942–948, 2015.

A. G. d. G. Matthews, J. Hensman, R. Turner, and Z. Ghahramani. On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. In *Artificial Intelligence and Statistics (AISTATS)*, pages 231–239, 2016.

G. J. McLachlan and K. E. Basford. *Mixture models: Inference and applications to clustering*, volume 38. M. Dekker New York, 1988.

M. R. Mehl, J. W. Pennebaker, D. M. Crow, J. Dabbs, and J. H. Price. The electronically activated recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers*, 33(4):517–523, 2001.

G. Miller. The Smartphone Psychology Manifesto. *Perspectives on Psychological Science*, 7(3): 221–237, 2012.

P. Moreno-Muñoz, D. Ramírez, and A. Artés-Rodríguez. Change-point detection on hierarchical circadian models. *arXiv preprint arXiv:1809.04197*, 2018.

P. Moreno-Muñoz, D. Ramírez, and A. Artés-Rodríguez. Change-point detection in hierarchical circadian models. *Pattern Recognition*, 2020.

P. Moreno-Muñoz, A. Artés-Rodríguez, and M. A. Álvarez. Heterogeneous multi-output Gaussian process prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6711–6720, 2018.

P. Moreno-Muñoz, A. Artés-Rodríguez, and M. A. Álvarez. Continual multi-task Gaussian processes. *arXiv preprint arXiv:1911.00002*, 2019.

P. Moreno-Muñoz, A. Artés-Rodríguez, and M. A. Álvarez. Recyclable Gaussian processes. *arXiv preprint arXiv:2010.02554*, 2020a.

P. Moreno-Muñoz, D. Ramírez, and A. Artés-Rodríguez. Continual learning for infinite hierarchical change-point detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3582–3586, 2020b.

P. Moreno-Muñoz, L. Romero-Medrano, Á. Moreno, J. Herrera-López, E. Baca-García, and A. Artés-Rodríguez. Passive detection of behavioral shifts for suicide attempt prevention. *Machine Learning for Mobile Health Workshop at NeurIPS*, 2020c.

K. P. Murphy. *Machine learning: A probabilistic perspective*. MIT Press, 2012.

A. Nazabal, P. Garcia-Moreno, A. Artes-Rodriguez, and Z. Ghahramani. Human activity recognition by combining a small number of classifiers. *IEEE Journal of Biomedical and Health Informatics*, 20(5):1342–1351, 2015.

A. Nazábal, P. Garcia-Moreno, A. Artés-Rodríguez, and Z. Ghahramani. Human activity recognition by combining a small number of classifiers. *IEEE Journal of Biomedical and Health Informatics*, 20(5):1342–1351, 2016.

A. Nazabal, P. M. Olmos, Z. Ghahramani, and I. Valera. Handling incomplete heterogeneous data using VAEs. *Pattern Recognition*, page 107501, 2020.

J. W. Ng and M. P. Deisenroth. Hierarchical mixture-of-experts model for large-scale Gaussian process regression. *arXiv preprint arXiv:1412.3078*, 2014.

C. V. Nguyen, T. D. Bui, Y. Li, and R. E. Turner. Online variational Bayesian inference: Algorithms for sparse Gaussian processes and theoretical bounds. In *Time Series Workshop @ ICML*, 2017.

C. V. Nguyen, Y. Li, T. D. Bui, and R. E. Turner. Variational continual learning. In *International Conference on Learning Representations (ICLR)*, 2018.

D. Nguyen-Tuong, J. R. Peters, and M. Seeger. Local Gaussian process regression for real time online model learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1193–1200, 2008.

J. C. Nunez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Velez. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. *Pattern Recognition*, 76:80–94, 2018.

Office of the Surgeon General (US) et al. 2012 National strategy for suicide prevention: Goals and objectives for action: a report of the US surgeon general and of the national action alliance for suicide prevention. 2012.

A. O'Hagan. Some bayesian numerical analysis. *Bayesian Statistics*, 4:345–363, 1992.

J.-P. Onnela and S. L. Rauch. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology*, 41(7):1691, 2016.

M. Opper and C. Archambeau. The variational gaussian approximation revisited. *Neural computation*, 21(3):786–792, 2009.

P. Orbanz and Y. W. Teh. Bayesian nonparametric models. *Encyclopedia of Machine Learning*, pages 81–89, 2010.

V. Osmani. Smartphones in Mental Health: Detecting Depressive and Manic Episodes. *IEEE Pervasive Computing*, 14(3):10–13, 2015.

C. Paciorek and M. Schervish. Nonstationary covariance functions for Gaussian process regression. *Advances in Neural Information Processing Systems (NeurIPS)*, 16:273–280, 2003.

G. Parra and F. Tobar. Spectral mixture kernels for multi-output Gaussian processes. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

I. Peis, P. M. Olmos, C. Vera-Varela, M. L. Barrigón, P. Courtet, E. Baca-García, and A. Artés-Rodríguez. Deep sequential models for suicidal ideation from multiple source data. *IEEE Journal of Biomedical and Health Informatics*, 23(6):2286–2293, 2019.

A. Pentland and A. Liu. Modeling and Prediction of Human Behavior. *Neural Computation*, 11(1): 229–242, 1999.

D. Peterson, P. Kanani, and V. J. Marathe. Private federated learning with domain adaptation. *Workshop on Federated Learning for Data Privacy and Confidentiality @ NeurIPS*, 2019.

J. Pitman. Combinatorial stochastic processes. Technical report, Dept. Statistics, UC Berkeley., 2002.

M. F. Pradier and F. Perez-Cruz. Infinite mixture of global Gaussian processes. In *Workshop in Bayesian Nonparametrics @ NIPS*, 2018.

C. E. Rasmussen and Z. Ghahramani. Infinite mixtures of Gaussian process experts. In *Advances in Neural Information Processing Systems (NIPS)*, pages 881–888, 2002.

C. E. Rasmussen and C. K. Williams. *Gaussian Processes for Machine Learning*, volume 2. MIT press, 2006.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and variational inference in deep latent Gaussian models.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

J. Robinson, G. Cox, E. Bailey, S. Hetrick, M. Rodrigues, S. Fisher, and H. Herrman. Social media and suicide prevention: A systematic review. *Early Intervention in Psychiatry*, 10(2):103–121, 2016.

L. Romero-Medrano, P. Moreno-Muñoz, and A. Artés-Rodríguez. Multinomial sampling for hierarchical change-point detection. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2020.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

F. J. Ruiz and M. K. Titsias. A contrastive divergence for combining variational inference and MCMC. In *International Conference on Machine Learning (ICML)*, pages 5537–5545, 2019.

F. J. Ruiz, M. K. Titsias, A. B. Dieng, and D. M. Blei. Augment and reduce: Stochastic inference for large categorical distributions. In *International Conference on Machine Learning (ICML)*, 2018.

Y. Saatçi, R. D. Turner, and C. E. Rasmussen. Gaussian process change point models. In *International Conference on Machine Learning (ICML)*, 2010.

B. Saha, T. Nguyen, D. Phung, and S. Venkatesh. A Framework for Classifying Online Mental Health-Related Communities With an Interest in Depression. *IEEE Journal of Biomedical and Health Informatics*, 20(4):1008–1015, July 2016.

A. Sano, A. J. Phillips, Z. Y. Amy, A. W. McHill, S. Taylor, N. Jaques, C. A. Czeisler, E. B. Klerman, and R. W. Picard. Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones. In *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 1–6. IEEE, 2015.

A. D. Saul, J. Hensman, A. Vehtari, and N. D. Lawrence. Chained Gaussian processes. In *Artificial Intelligence and Statistics (AISTATS)*, pages 1431–1440, 2016.

A. J. Scott and M. Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, (30):507–512, 1974.

M. Seeger. PAC-Bayesian generalisation error bounds for Gaussian process classification. *Journal of Machine Learning Research (JMLR)*, 3(Oct):233–269, 2002.

M. Seeger. Bayesian Gaussian process models: PAC-Bayesian generalisation error bounds and sparse approximations. *PhD Thesis, University of Edinburgh*, 2003.

A. Shah, A. Wilson, and Z. Ghahramani. Student-t processes as alternatives to Gaussian processes. In *Artificial Intelligence and Statistics (AISTATS)*, pages 877–885, 2014.

G. E. Simon, E. Johnson, J. M. Lawrence, R. C. Rossom, B. Ahmedani, F. L. Lynch, A. Beck, B. Waitzfelder, R. Ziebell, R. B. Penfold, et al. Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. *American Journal of Psychiatry*, 175 (10):951–960, 2018.

G. Skolidis and G. Sanguinetti. Bayesian multitask classification with Gaussian process priors. *IEEE Transactions on Neural Networks*, 22(12), 2011.

V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4424–4434, 2017.

E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1257–1264, 2006.

H. Soleimani, J. Hensman, and S. Saria. Scalable joint models for reliable uncertainty-aware event prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(8): 1948–1963, 2018.

A. Solin and S. Särkkä. Explicit link between periodic covariance functions and state space models. In *Artificial Intelligence and Statistics*, 2014.

A. Solin, J. Hensman, and R. E. Turner. Infinite-horizon Gaussian processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3486–3495, 2018.

C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

A. Stone, S. Shiffman, A. Atienza, and L. Nebeling. *The science of real-time data capture: Self-reports in health research*. Oxford University Press, 2007.

S. A. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard. Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Transactions on Affective Computing*, 2017.

Y. Teh, M. Seeger, and M. Jordan. Semiparametric latent factor models. In *Artificial Intelligence and Statistics (AISTATS)*, pages 333–340, 2005.

M. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics (AISTATS)*, pages 567–574, 2009a.

M. K. Titsias. Variational model selection for sparse Gaussian process regression. *Technical Report, University of Manchester*, 2009b.

M. K. Titsias, J. Schwarz, A. G. de G. Matthews, R. Pascanu, and Y. W. Teh. Functional regularisation for continual learning with Gaussian processes. In *International Conference on Learning Representations (ICLR)*, 2020.

J. Torous, P. Staples, M. Shanahan, C. Lin, P. Peck, M. Keshavan, and J.-P. Onnela. Utilizing a personal smartphone custom app to assess the patient health questionnaire-9 (PHQ-9) depressive symptoms in patients with major depressive disorder. *JMIR Mental Health*, 2(1):e8, 2015.

J. Torous, J. Onnela, and M. Keshavan. New dimensions and new tools to realize the potential of RDoC: Digital phenotyping via smartphones and connected devices. *Translational Psychiatry*, 7 (3), 2017.

J. Torous, M. E. Larsen, C. Depp, T. D. Cosco, I. Barnett, M. K. Nock, and J. Firth. Smartphones, sensors, and machine learning to advance real-time prediction and interventions for suicide prevention: A review of current progress and next steps. *Current Psychiatry Reports*, 20(7):51, 2018a.

J. Torous, H. Wisniewski, G. Liu, and M. Keshavan. Mental health mobile phone app usage, concerns, and benefits among psychiatric outpatients: Comparative survey study. *JMIR Mental Health*, 5(4):e11715, 2018b.

V. Tresp. A Bayesian committee machine. *Neural Computation*, 12(11):2719–2741, 2000.

R. Turner, Y. Saatci, and C. E. Rasmussen. Adaptive sequential Bayesian change point detection. In *Workshop in Advances in Neural Information Processing Systems (NIPS)*, 2009.

R. D. Turner, S. Bottone, and C. J. Stanek. Online variational approximations to non-exponential family change point models: with application to radar tracking. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

K. R. Ulrich, D. E. Carlson, K. Dzirasa, and L. Carin. GP kernels for cross-spectrum analysis. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

I. Valera and Z. Ghahramani. Automatic discovery of the statistical types of variables in a dataset. In *International Conference on Machine Learning (ICML)*, pages 3521–3529, 2017.

I. Valera, M. F. Pradier, M. Lomeli, and Z. Ghahramani. General latent feature models for heterogeneous datasets. *arXiv preprint arXiv:1706.03779*, 2017.

I. Valera, M. F. Pradier, M. Lomeli, and Z. Ghahramani. General latent feature models for heterogeneous datasets. *Journal of Machine Learning Research (JMLR)*, 21(100):1–49, 2020.

M. Van der Wilk, C. E. Rasmussen, and J. Hensman. Convolutional Gaussian processes. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2849–2858, 2017.

J. Vanhatalo, J. Riihimäki, J. Hartikainen, P. Jylänki, V. Tolvanen, and A. Vehtari. GPstuff: Bayesian modeling with Gaussian processes. *Journal of Machine Learning Research (JMLR)*, 14 (1):1175–1179, 2013.

D. Velychko, B. Knopp, and D. Endres. Making the coupled Gaussian process dynamical model modular and scalable with variational approximations. *Entropy*, 20(10):724, 2018.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

K. Wang, G. Pleiss, J. Gardner, S. Tyree, K. Q. Weinberger, and A. G. Wilson. Exact Gaussian processes on a million data points. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 14648–14659, 2019.

C. K. Wikle. A kernel-based spectral model for non-Gaussian spatio-temporal processes. *Statistical Modelling*, 2(4):299–314, 2002.

C. K. Williams. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. In *Learning in Graphical Models*, pages 599–621. Springer, 1998.

A. G. Wilson and Z. Ghahramani. Generalised Wishart processes. In *Uncertainty in Artificial Intelligence (UAI)*, 2011.

World Health Organization. Mental disorders key facts. 2019.

World Health Organization et al. The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. *Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire*, 67(30):227–227, 1992.

Y. Xie, J. Huang, and R. Willett. Change-point detection for high-dimensional time series with missing data. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):12–27, 2013.

X. Xuan and K. Murphy. Modeling changing dependency structure in multivariate time series. In *International Conference on Machine Learning (ICML)*, 2007.

Y. Yamashita, M. Onodera, K. Shimoda, and Y. Tobe. Visualizing health with emotion polarity history using voice. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, pages 1210–1213, 2019.

L. Yang, K. Wang, and L. S. Mihaylova. Online sparse multi-output Gaussian process regression and learning. *IEEE Transactions on Signal and Information Processing over Networks*, 5(2):258–272, 2018.

M. D. Zeiler. ADADELTA: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

M. M. Zhang, B. Dumitrascu, S. A. Williamson, and B. E. Engelhardt. Sequential Gaussian processes for online learning of nonstationary functions. *arXiv preprint arXiv:1905.10003*, 2019.

C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560, 1997.