FAMAF Facultad de Matemática, Astronomía, Física y Computación

UNC Universidad Nacional de Córdoba

Asimilación de datos por ensambles y tratamiento de errores: aplicaciones en modelos epidemiológicos

por

Tadeo Javier COCUCCI

Presentado ante la Facultad de Matemática, Astronomía, Física y Computación como parte de los requerimientos para la obtención del grado de Doctor en Ciencias de la Computación de la

UNIVERSIDAD NACIONAL DE CÓRDOBA

Julio, 2022

Director: Dr. Manuel PULIDO

Tribunal Especial:

Titulares:

Dr. Damián FERNÁNDEZ FERREYRA (CONICET, FaMAF-UNC)
Dr. Claudio DELRIEUX (CONICET, ICIC-UNS)
Dr. Jorge SÁNCHEZ (CONICET)

Suplentes:

Dra. Milagros TERUEL (Microsoft, FaMAF-UNC)
Dr. Germán TORRES (CONICET, FACENA-UNNE)

Este trabajo se distribuye bajo una Licencia Creative Commons Atribución-NoComercial-CompartirIgual 4.0 Internacional

Resumen

En este trabajo abordamos uno de los desafíos principales dentro de la disciplina de la asimilación de datos: la especificación de las incertezas inherentes al modelo y a las observaciones en los sistemas parcialmente observados donde se aplican típicamente las técnicas de asimilación. Para ello presentamos el problema de asimilación de datos desde una perspectiva Bayesiana y los métodos más relevantes, con especial énfasis en las técnicas basadas en ensambles. Además introducimos el problema de la especificación de los errores de modelo y observacionales y como esta afecta en la performance de la asimilación de datos. También exponemos el panorama de estrategias que se suelen utilizar para estimarlos. Proponemos un método de inferencia de estos errores basado en el algoritmo EM (expectation-maximization). A diferencia de las implementaciones clásicas del EM, nuestra versión incorpora observaciones una a una. Esto la hace especialmente adecuada para contextos de asimilación de datos secuencial. Evaluamos el método propuesto de manera extensiva mediante experimentos gemelos.

El modelado epidemiológico es interpretable dentro del marco teórico en el que es posible utilizar técnicas de asimilación de datos. Debido a esto y su importancia en materia de salud pública, tomamos a la epidemiología como campo de aplicación de las metodologías desarrolladas y estudiamos la factibilidad de su aplicación en modelos epidemiológicos compartimentales clásicos. Aún cuando los modelos compartimentales expresan a grandes rasgos la propagación de enfermedades en una población, estos carecen de aspectos que son de importancia para el estudio de la dinámica de contagios que permiten definir políticas de confinamiento, trazado de contactos, diseño de campañas de vacunación, etc. Los modelos basados en agentes diseñan el escenario desde la descripción en detalle de cada individuo lo que les da gran flexibilidad y expresividad. Debido a esto, abordamos la posibilidad de utilizar métodos de asimilación de datos por ensambles en modelos epidemiológicos basados en agentes. Este es un enfoque novedoso al problema ya que en general se suele tratar con modelos dinámicos basados en ecuaciones diferenciales. Por otro lado, constituye un desafío puesto que las técnicas de asimilación de datos están típicamente diseñadas para ser utilizadas en sistemas en los que las variables de estado se modelan explícitamente. Esto no es así en los modelos basados en agentes, en los que el estado emerge como el comportamiento colectivo de las interacciones en la escala de los individuos. Con estos aportes mostramos el potencial de la aplicación de técnicas de asimilación de datos para estimar parámetros, errores y variables de estado en sistemas epidemiológicos.

Abstract

This work deals with one of the main challenges in the data assimilation framework: the specification of the inherent uncertainties of the model and measurements in partially observed systems where data assimilation techniques are typically applied. To do this, we present the data assimilation problem from a Bayesian perspective and the most relevant methods, with special emphasis on ensemble-based techniques. We also address the issue of the specification of model and observational error and how these affect on data assimilation performance. We give an overview of the strategies used to estimate them. We propose an inference method for these errors based on the EM algorithm. Unlike classic implementations of EM, our version processes observations one by one. This makes it specially adequate in sequential data assimilation contexts. We extensively evaluate the method with twin experiments.

Epidemiological modeling is interpretable within the theoretical framework in which data assimilation is applied. Because of this, and its importance regarding public health, we take epidemiological modeling as the field of application for the developed techniques, and we study the feasibility of their use with classical compartmental models. Even though compartmental models give an overall characterization of the propagation of a disease within a population, these lack features which are of importance to study contagion dynamics which allow defining confinement policies, contact tracing, vaccination campaigns schedules, etc. Agent-based models design scenarios by describing in detail each individual which grants them great expressiveness and flexibility. Because of this we approach the possibility of utilizing ensemble-based data assimilation for agent-based models. This is a novel perspective to the problem since usually, in this context, dynamic models based on differential equations are used. On the other hand, it poses a challenge because data assimilation techniques are typically designed for systems in which state variables are modeled explicitly. This is not the case for agent-based models, in which state emerges from the collective behaviour of the individual-scale interactions. With these contributions we showcase the potential of applying data assimilation to estimate parameters, errors and state variables within epidemiological systems.

Tabla de contenidos

Capítulo 1

Introducción

En esta tesis exploramos algunos de los desafíos asociados a la aplicación de técnicas de asimilación de datos basadas en ensambles sobre modelos epidemiológicos. Abordamos el problema de la especificación de la incerteza inherente al modelo y las observaciones tanto en un marco general de modelos parcialmente observados como para el caso específico de modelos epidemiológicos basados en ecuaciones diferenciales. Además estudiamos el potencial de utilizar técnicas de asimilación de datos en modelos basados en agentes.

Muchos sistemas complejos suelen ser de alta dimensionalidad. Por ejemplo, sistemas sociales, geofísicos, ecológicos, etc. suelen tener esta característica. Además, debido a su tamaño y complejidad suele ser difícil tomar mediciones de su estado. La información que se consigue a través de observaciones es entonces parcial y con errores de significativos. En este contexto se requiere incluir información basada en el conocimiento del sistema. Para esto, se pueden considerar modelos matemáticos con el fin de generar pronósticos y predecir su comportamiento. Estas dos fuentes de información están en principio incomunicadas. La asimilación de datos comprende un conjunto de técnicas estadísticas que se utilizan para combinar estas fuentes de información sobre el estado del sistema: pronósticos provenientes de modelos matemáticos y observaciones (Kalnay, 2003). Ambas fuentes de información son propensas a errores. El error de modelo cuantifica nuestro conocimiento limitado de la dinámica del sistema, aproximaciones y errores numéricos. El error observacional incluye la incerteza propia de los instrumentos de medición y el error de representatividad que involucra como se relacionan las observaciones con el estado del sistema. Esto se describe con mayor detalle en la Sección 3.1. La asimilación de datos apunta a encontrar una combinación ponderada entre estas fuentes de información, de manera que si sabemos que la incerteza del modelo es menor que la de los datos, la estimación resultante será más fiel al modelo y si por el contrario, las observaciones tienen menos error que el pronóstico la estimación será más próxima a los datos.

Una de las aplicaciones pioneras de la asimilacion de datos es la predicción numérica meteorológica. La meteorología es un sistema complejo por antonomasia, con numerosos procesos dinámicos de diferentes escalas que interactuan entre sí. Por las características de la atmósfera y la tierra las observaciones son escasas. Esto motivó originalmente los desarrollos de la asimilación variacional de datos (Talagrand y Courtier, 1987). Este enfoque permite reformular el problema como la minimización de una función de costo que penaliza a las fuentes de información de mayor incerteza. En el área de la predicción meteorológica numérica se cuenta con modelos matemáticos y computacionales muy complejos y de alta dimensionalidad que describen la evolución de los procesos físicos e informan sobre diversas variables de estado (por ejemplo, velocidad, temperatura o presión) en diferentes puntos de una grilla espacial potencialmente muy grande ($10^7 - 10^8$ dimensiones).

Estos modelos se basan en leyes físicas y permiten obtener pronósticos: por ejemplo, las ecuaciones de Navier-Stokes que expresan la conservación del momento y de la masa en fluidos. Por otro lado, se tiene otra fuente de información sobre el mismo sistema que consta de las observaciones de diversos instrumentos en estaciones meteorológicas o provenientes de satélites.

El filtro de Kalman (Kalman y Bucy, 1961; Kalman, 1960) ocupa un lugar central dentro de las técnicas de asimilación de datos pues es una metodología sencilla que ha sentado las bases para métodos más sofisticados. Este tipo de filtro lineal encontró aplicaciones relevantes en la determinación de órbitas satelitales, navegación de submarinos y aeronaves e incluso de misiones espaciales como la Apollo (Jazwinski, 1970). En esta clase de aplicaciones típicamente tenemos que, con el fin de estimar la posición y velocidad, se utiliza como modelo a las ecuaciones físicas de movimiento mientras que las observaciones provienen de los instrumentos de navegación. Una gran parte del desarrollo de técnicas de asimilación de datos proviene sin embargo del área de las geociencias donde se presentan otro tipo de desafíos como la alta dimensionalidad de los sistemas, observaciones menos precisas y modelos caóticos altamente no lineales. La alta dimensionalidad de estos sistemas hace imposible la representación de la matriz de covarianza de la prediccion requerida por el filtro de Kalman (de dimensiones inmensas, $10^7 \times 10^7$ o mayores). Esto motivó el desarrollo del filtro de Kalman por ensambles (Evensen, 1994), el cual toma la idea original de Kalman incorpora la idea de representar distribuciones mediante muestras lo cual permite adaptar el problema a situaciones de no linealidad. Así además se evita la representación explícita de la matriz de covarianza de las predicciones lo que permite la aplicación a sistemas de mayores dimensiones. La familia de métodos por ensambles pudo competir con los métodos variacionales (3D-VAR y 4D-VAR) que son utilizados en grandes centros meteorológicos (Kalnay et al., 2007).

Más allá que las técnicas de asimilación de datos fueron desarrolladas inicialmente para resolver problemas asociados a la predicción numérica meteorológica y para la aeronavegación espacial (Grewal y Andrews, 2010), actualmente el campo de aplicación es mucho más amplio. Se utilizan, por ejemplo, para la predicción sobre reservorios petrolíferos Aanonsen et al., 2009, oceanografía, detección de incendios forestales (Mandel et al., 2008), epidemiología (Shaman y Karspeck, 2012), entre otras.

Desde el punto de vista conceptual, existen dentro de los métodos paramétricos no lineales de la asimilación de datos principalmente dos formulaciones: los filtros de Kalman por ensambles y los métodos variacionales. Por otro lado tenemos que, más recientemente, los filtros de partículas han cobrado relevancia. Estos están basados en la aplicación secuencial de Monte Carlo y habilitan la representación de distribuciones no Gaussianas. El conjunto conforma una gran familia de metodologías de creciente interés (Van Leeuwen et al., 2019). En general, los filtros de partículas no hacen suposiciones acerca de las distribuciones involucradas, por lo que en principio se pueden aplicar a cualquier variante del problema de asimilación de datos. Sin embargo, la representación de distribuciones totalmente generales mediante muestras requiere de un número muy grande de partículas lo que puede resultar privativo.

La variedad de metodologías de asimilación es vasta, en especial porque para cada técnica se tiene una plétora de variantes e implementaciones que se utlizan para mitigar los problemas que presenta cada aplicación. Así tenemos una familia de filtros de Kalman por ensambles, otra de métodos variacionales, otra de filtros de partículas, etc. Incluso, dentro de cada una de estas familias tenemos subgrupos de técnicas que comparten ciertas características o abordajes conceptuales. Más aún,

nuevos enfoques apuntan a recuperar lo mejor de cada metodología mediante híbridos: por ejemplo existen híbridos entre filtros de Kalman por ensambles y métodos variacionales (Hamill y Snyder, 2000) o con filtros de partículas (Frei y Künsch, 2013; Stordal et al., 2011). En Carrassi, Bocquet, Bertino et al., 2018 se puede encontrar un buen panorama del estado del arte en cuanto a metodologías de asimilación de datos.

Actualmente el creciente interés en *machine learning* y los grandes avances en el área en las últimas décadas motivó interés en el intercambio entre esta área y la asimilación de datos. En Abarbanel et al., 2018 se plantea una interpretación de las redes neuronales en el contexto de problemas de asimilación de datos dando una equivalencia entre agregar capas a la red neuronal con la resolución temporal en el problema de asimilación. En la asimilacion de datos variacional, el modelo adjunto es equivalente al método de *backpropagation* y en ambos casos se determina la sensibilidad de la función de costo a párametros de la red o variables del sistema. Además, se señala la utilidad de algunas herramientas de asimilación para su aplicación en *machine learning*, en particular el uso de *variational annealing* para encontrar mínimos globales en funciones de costo. Por otro lado, en Kovachki y Stuart, 2019 se propone el uso técnicas basadas en filtros de Kalman por ensambles para asistir el entrenamiento de modelos supervisados y semi-supervisados de manera que se evita el uso de gradientes. Potencialmente, esta metodología se podría utilizar para capturar la incerteza en una red neuronal entrenada usando la variabilidad del ensamble. Con estas estrategias se pueden entrenar y determinar los parametros de la red sin el requerimiento de *backpropagation*; sólo con la evolución hacia adelante de un ensamble de estados del modelo se determinan las correlaciones entre las variables y los parámetros.

Como mencionamos anteriormente, la asimilación de datos tiene en cuenta la incerteza del modelo que genera los pronósticos del sistema tanto como la proveniente de las observaciones. Una especificación errónea de estas cantidades puede causar una performance subóptima de la inferencia pero es habitual que estos errores sean difíciles de identificar. En general, en la práctica, se les da valores *a priori* a los errores asociados al modelo y las observaciones pero esto puede ser en detrimento del desempeño del sistema. Este problema constituye uno de los mayores desafíos en el área y se han desarrollado una variedad de métodos para proveer estimaciones para el error observacional y de modelo (Tandeo, Ailliot et al., 2020). Estos incluyen metodologías basadas en momentos estadísticos y otras que apuntan a la maximización de la verosimilitud. Entre estas últimas hacemos mención del algoritmo EM (Dempster et al., 1977), el cual puede ser implementado en el contexto de asimilación de datos con filtros de Kalman y en particular con filtros basados en ensambles (Tandeo, Pulido et al., 2015).

La implementación clásica del algoritmo EM toma un lote (o *batch*) de datos y los procesa a todos juntos para proveer estimaciones de los parámetros que codifican los errores. Esto trae una serie de inconvenientes a la hora de ser utilizado en sistemas de asimilación secuencial en los que los datos son producidos y procesados en tiempo real o casi real. El procesamiento por lotes es *offline* por lo que, por un lado, fuerza a almacenar todo el conjunto de observaciones y por otro, cuando ingresa una observación no es en principio adaptable a incorporarla por lo que el proceso debe reiniciarse. Esta situación ha despertado interés en técnicas de inferencia *online*, es decir que permitan actualizaciones con cada nueva observación y de manera que cada dato es procesado una única vez. En esta dirección se destaca la metodología desarrollada en Berry y Sauer, 2013 la cual es basada en innovaciones (esto, a grandes rasgos, significa que se basa en las diferencias entre pronósticos y

observaciones) y que está específicamente diseñada para contextos de asimilación de datos. Por otro lado, luego del trabajo seminal de Neal y Hinton, 1998 tenemos versiones *online* del EM por Andrieu y Doucet, 2003 y por Cappé (Cappé, 2009; Cappé, 2011). Estas sin embargo no están explícitamente diseñadas para ser acopladas a técnicas clásicas de asimilación de datos. Por el interés en metodologías de este tipo y su relevancia, en esta tesis planteamos el desarrollo de versiones *online* del EM que se adapta a métodos de asimilación por ensambles, tanto filtros de partículas como el filtro de Kalman por ensambles. Además evaluamos experimentalmente la metodología en escenarios típicos de asimilación de datos.

El marco en el que se suele aplicar la asimilación de datos es el de un sistema que evoluciona temporalmente y es parcialmente observado. Estos pueden interpretarse teóricamente como modelos de Markov escondidos o *state-space models*. Los sistemas epidemiológicos tienen estas características y por lo tanto la utilización de herramientas de asimilación de datos en estos sistemas puede ser provechosa. Desde comienzos del siglo pasado se han utilizado modelos matemáticos para estudiar las dinámicas de transmisión de enfermedades. Los populares modelos compartimentales basados en ecuaciones diferenciales (se suele señalar a Kermack et al., 1927 como el principal antecedente) han sido adaptados a una gran diversidad de escenarios epidemiológicos permitiendo predicción de picos epidémicos, simulación de medidas de prevención y mitigación, estimación de parámetros, etc. Estos se basan en considerar una subpoblación dividida en compartimentos de acuerdo a su estado epidemiológico. Por ejemplo, un modelo sencillo de este tipo podría considerar a las subpoblaciones de susceptibles, infectados y recuperados. En general se utiliza una ecuación diferencial para representar la evolución temporal de cada uno de estos compartimentos.

Los ejemplos de aplicación de asimilación de datos sobre estos modelos epidemiológicos no son tan abundantes como en otras áreas. El trabajo de Ionides et al., 2006 es relevante en este aspecto puesto que utiliza filtros de partículas sobre un modelo de cólera. Además introduce una técnica de inferencia para sistemas no lineales que permite la estimación de parámetros del modelo. Por otro lado Shaman y Karspeck, 2012 usa el filtro de Kalman por ensambles acoplado a un modelo de gripe para predecir picos epidémicos en Nueva York. Luego, con el advenimiento de la pandemia de COVID-19 este tipo de abordaje se hizo algo más popular (por ejemplo Evensen et al., 2020; Li, Pei et al., 2020). A pesar de que la asimilación de datos en estos sistemas comenzó a ser más común, la estimación de errores observacionales y de modelo en estos trabajos es infrecuente. Los modelos epidemiológicos suelen simplificar numerosas características dentro de la complejidad inherente a sistemas de este tipo, por lo que el error de modelo es inevitable. Por otro lado las observaciones sobre estos sistemas suelen constar de reportes de instituciones de salud y pueden incluir una diversidad de errores como subreportes, diagnósticos incorrectos, heterogeneidad en el monitoreo, etc. Esto motivó que tomemos un modelo epidemiológico compartimental sencillo para COVID-19 como ejemplo de aplicación de nuestro algoritmo EM *online* con el objetivo de estudiar el efecto de estos errores en sistemas epidemiológicos y su respuesta al ser acoplados a métodos de inferencia de este tipo.

Los modelos epidemiológicos compartimentales están típicamente representados mediante ecuaciones diferenciales y las técnicas de asimilación de datos son en general aplicadas a modelos dinámicos con esta característica. Sin embargo, se han popularizado, en parte a las posibilidades que abre el aumento del poder de cómputo, los modelos basados en agentes. Estos, en lugar de tomar las variables de interés y representar su comportamiento mediante ecuaciones, simulan una población

de individuos a través de un conjunto de reglas de interacción. El comportamiento global del sistema es entonces el resultado del comportamiento colectivo de la totalidad de los agentes. Este paradigma provee gran adaptabilidad y expresividad a estos modelos ya que permiten representar, mediante reglas simples de interrelación entre individuos, situaciones que pueden ser muy difíciles de reproducir mediante ecuaciones diferenciales. Los modelos basados en agentes pueden constituir una representación de alguna manera más directa para sistemas sociales. Es común que estos modelos requieran de una gran cantidad de parámetros debido a la cantidad de detalles que se deben tener en cuenta. Esto, sumado a la creciente popularidad y complejización de este tipo de modelos, ha causado que cobre relevancia el problema de estimar los parámetros para lograr una calibración adecuada. Además, para mejorar las capacidades de los modelos basados en agentes, es importante realizar tareas de inferencia estadística de manera de utilizar datos observacionales de los fenómenos que se modelan y así mejorar calidad de la representación. Esto permitiría obtener evoluciones temporales de la población de individuos acopladas a las restricciones estadisticas de los datos.

La asimilación de datos ha encontrado en el modelado epidemiológico más tradicional un campo de aplicación con resultados favorables y por lo tanto resulta prometedora la idea de extender su aplicabilidad a modelos basados en agentes. Con la notable excepción del trabajo de Ward et al., 2016, en el que trabaja con un modelo de movimiento de peatones, no existen en nuestro conocimiento ejemplos de la utilización de técnicas de asimilación de datos en modelos de agentes. Esto nos motivó a seguir esta línea de investigación para lo cual desarrollamos una metodología general para utilizar técnicas de asimilación por ensambles con modelos de agentes. Como aplicación diseñamos un modelo de COVID-19 para el cual evaluamos experimentalmente el método. El diseño de los experimentos busca dar cuenta de algunas de las dificultades comunes en la inferencia de sistemas de propagación de COVID-19, como la estimación de la cantidad de asintomáticos, así como desafíos típicos de asimilación de datos tales como dar cuenta del error proveniente de un modelo mal especificado.

El surgimiento del COVID-19 dio lugar a que, entendiblemente, el interés de la comunidad científica se concentre en esta enfermedad y que proliferen modelos y métodos de inferencia para intentar comprender mejor este fenómeno, intentar dar predicciones y asesorar instituciones de salud y a tomadores de decisiones con el propósito de minimizar los daños de la crisis sanitaria global que significó la pandemia. Debido a esto y también por la gran disponibilidad de datos accesibles por el monitoreo de contagios, decidimos usar al COVID-19 como aplicación principal de los métodos desarrollados en la tesis.

En el Capítulo 2 formulamos el problema de la asimilación de datos desde una perspectiva Bayesiana e introducimos las técnicas más relevantes para el desarrollo de nuestro trabajo. En el Capítulo 3 introducimos el problema de tratamiento de errores, discutimos algunas de las metodologías más conocidas y finalmente presentamos nuestro algoritmo EM *online* con su deducción teórica y una evaluación experimental de su desempeño. En el Capítulo 4 introducimos los elementos básicos del modelado epidemiológico y los antecedentes de inferencia estadística basada en asimilación de datos en este tipo de escenario. Luego mostramos los resultados de la aplicación del EM *online* en un modelo epidemiológico compartimental. En el Capítulo 5 damos un marco general de aplicación de asimilación de datos basada en ensambles para modelos basados en agentes y hacemos una evaluación experimental de la metodología sobre un modelo que desarrollamos en base a las características

epidemiológicas del COVID-19. Finalmente, damos nuestras conclusiones y discutimos potenciales líneas de investigación en el Capítulo 6.

Capítulo 2

Asimilación de datos

2.1 Asimilación de datos como un problema de inferencia Bayesiana

La asimilación de datos busca hacer inferencia sobre variables de estado $\mathbf{x} \in \mathbb{R}^{N_x}$ incorporando información observacional $\mathbf{y} \in \mathbb{R}^{N_y}$. Si consideramos que estas son variables aleatorias, el problema de incorporar la información observacional a las variables de estado se puede pensar como encontrar la distribución condicional $p(\mathbf{x}|\mathbf{y})$, llamada también distribución *a posteriori*. En este tipo de escenarios, es natural aplicar la regla de Bayes para obtener

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}$$

donde $p(\mathbf{y}|\mathbf{x})$ es la verosimilitud, $p(\mathbf{x})$ es la distribución *a priori* de las variables de estado, y va a estar determinada por nuestro modelo de pronóstico, mientras que $p(\mathbf{y})$ puede ser vista como una constante de normalización pues no depende de $\mathbf{x}$. La verosimilitud se interpreta como una función de $\mathbf{x}$ y nos informa cuan factible es que la observación $\mathbf{y}$ haya sido producida por el estado $\mathbf{x}$. La verosimilitud de $\mathbf{x}$ habiéndose observado $\mathbf{y}$ se suele denotar como $\mathcal{L}(\mathbf{x};\mathbf{y})$ para enfatizar que no es una densidad de probabilidad y que es función de $\mathbf{x}$. El modelo observacional, que es la representación de cómo se obtiene un dato desde las variables de estado va a determinar a la verosimilitud.

2.1.1 State-space model

Si consideramos que tenemos un proceso en el que las variables de estado evolucionan temporalmente, podemos denotar a un conjunto de realizaciones del proceso como $\mathbf{x}_{0:t} = \mathbf{x}_0, ..., \mathbf{x}_t$ y, de manera similar, a las observaciones sobre ese proceso como $\mathbf{y}_{1:t} = \mathbf{y}_1, ..., \mathbf{y}_t$. Comúnmente, el proceso $\mathbf{x}_{0:t}$ es la discretización de un proceso en tiempo continuo y esta discretización suele estar determinada por la disponibilidad de las observaciones que normalmente están equiespaciadas temporalmente. Las variables de estado evolucionan del tiempo t al $t+1$ a través de un modelo $\mathcal{M}_t$ y a su vez, el modelo observacional $\mathcal{H}_t$ es el que representa como se obtiene la observación $\mathbf{y}_t$ del estado $\mathbf{x}_t$:

$$\mathbf{x}_t = \mathcal{M}_t(\mathbf{x}_{t-1}, \boldsymbol{\eta}_t), \tag{2.1}$$

$$\mathbf{y}_t = \mathcal{H}_t(\mathbf{x}_t, \boldsymbol{\nu}_t). \tag{2.2}$$

En estas ecuaciones introducimos $\boldsymbol{\eta}_t$ y $\boldsymbol{\nu}_t$ como las componentes estocásticas que dan cuenta del error de modelo y observacional respectivamente.

Notemos además que la Ecuación 2.1 determina una probabilidad de transición $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ y la Ecuación 2.2 define una verosimilitud observacional $\mathcal{L}_t(\mathbf{x}_t;\mathbf{y}_t) = p(\mathbf{y}_t|\mathbf{x}_t)$. Es una convención en asimilación de datos considerar a las variables de estado indexadas desde el 0 y a las observaciones desde el 1. De esta manera se asume que $\mathbf{x}_0$, el *prior* de las variables de estado antes de la primera observación, no es observado. Si además suponemos que el estado inicial responde a una distribución, $p(\mathbf{x}_0)$, podemos plantear al problema de la siguiente manera:

$$\mathbf{x}_0 \sim p(\mathbf{x}_0) \tag{2.3}$$
$$\mathbf{x}_t|\mathbf{x}_{t-1} \sim p(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{2.4}$$
$$\mathbf{y}_t|\mathbf{x}_t \sim p(\mathbf{y}_t|\mathbf{x}_t) \tag{2.5}$$

Si no hacemos suposiciones sobre el modelo $\mathcal{M}_t$ es difícil saber cual será el efecto de la evolución temporal del sistema, 2.4, sobre la distribución variables de estado, incluso en el caso que la distribución inicial de $\mathbf{x}_0$ sea sencilla. Por ejemplo, modelos no lineales de baja dimensionalidad pueden llevar a que una distribución inicial Gaussiana resulte multimodal al ser evolucionada hacia adelante.

Ejemplo Lorenz-63 Tomamos como ejemplo, el modelo Lorenz-63 de 3 dimensiones (Lorenz, 1963), determinado por las ecuaciones,

$$\begin{cases} \dfrac{\partial x}{\partial t} = \sigma(x-y) \\ \dfrac{\partial y}{\partial t} = x(\rho - x) - y \\ \dfrac{\partial z}{\partial t} = xy - \beta z \end{cases} \tag{2.6}$$

y consideramos que $\mathcal{M}_t$ consiste en la integración temporal de este modelo. Las no linealidades que tiene son suficientes para causar que se pierda la Gaussianidad de manera muy evidente. Para mostrarlo, integramos 2.6 desde 100 puntos muestrales tomados de una distribución normal durante 100 pasos de tiempo de longitud 0.01. En la Figura 2.1 se muestran las trayectorias proyectadas sobre el plano xy. Se puede ver que la distribución de los puntos luego de la aplicación del modelo es claramente bimodal. Este ejemplo exagera el efecto puesto que permite una integración muy larga del modelo sin hacer ninguna restricción, pero si observamos la manera en que se separan las trayectorias se hace evidente que la Gaussianidad se puede perder rápidamente cuando el modelo $\mathcal{M}$ es no lineal.

2.1.2 Modelo de Markov escondido

El modelo propuesto por las Ecuaciones 2.1 y 2.2 constituye un modelo de Markov escondido. En este tipo de representaciones, se tiene que las variables de estado $\{\mathbf{x}_t\}_{t\geq 0}$ son una cadena de Markov la cual no es directamente observable. Debido a esto, a las variables $\mathbf{x}_t$ se las llama variables escondidas o latentes. A su vez, la información sobre esta cadena proviene de un proceso que sí es observable $\{\mathbf{y}_t\}_{t\geq 1}$. La Figura 2.2 representa este tipo de configuración. En este esquema, buscamos inferir sobre el estado escondido utilizando la información del proceso observable. Más formalmente, las propiedades que definen a un proceso de Markov escondido son:

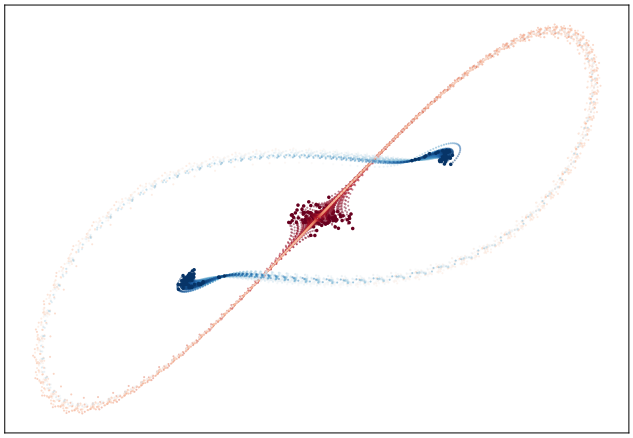

Figura 2.1: Trayectorias del modelo Lorenz-63. Los puntos iniciales (rojos) y finales (azules) están representados con un mayor tamaño que los intermedios.

1. **El proceso $\{\mathbf{x}_t\}_{t\geq 0}$ es una cadena de Markov** lo que significa que el proceso "no tiene memoria", es decir que $p(\mathbf{x}_t|\mathbf{x}_{0:t-1}) = p(\mathbf{x}_t|\mathbf{x}_{t-1})$: si el estado a tiempo $t-1$ está determinado, x_t depende sólo de este y no de estados anteriores. Esto permite escribir:

$$p(\mathbf{x}_{0:t}) = p(\mathbf{x}_0)\prod_{k=1}^{t} p(\mathbf{x}_k|\mathbf{x}_{k-1})$$

2. **Las observaciones son condicionalmente independientes** lo cual implica que $p(\mathbf{y}_t|\mathbf{x}_{0:t}) = p(\mathbf{y}_t|\mathbf{x}_t)$, es decir que la observación a tiempo t sólo depende del estado a tiempo t (y no de otros). Esto además resulta en que:

$$p(\mathbf{y}_{1:t}|\mathbf{x}_{0:t}) = \prod_{k=1}^{t} p(\mathbf{y}_k|\mathbf{x}_k)$$

El diagrama de la Figura 2.2 puede ser interpretado como un modelo gráfico, con lo cual las propiedades de Markovianidad e independencia condicional de las observaciones se deducen de manera automática de esta interpretación. Esto habilita además el uso de propiedades de modelos gráficos como la D-separación (ver Jordan, 1999).

Existen otros marcos teóricos para construir modelos probabilísticos para datos secuenciales que dan lugar a algoritmos de estimación de parámetros distintos de los que presentaremos en este trabajo. Por ejemplo, en el área del procesamiento del lenguaje natural se utilizan también modelos de Markov de máxima entropía (MEMMs) y *Conditional Random Fields* (CRFs) los cuales tienen propiedades estadísticas diferentes a los modelos de Markov escondidos pero que sin embargo pueden ser aplicados para tratar con problemas similares (Lafferty et al., 2001).

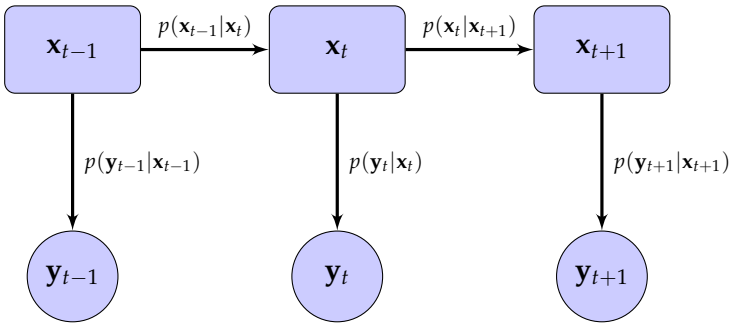

Figura 2.2: Esquematización de un modelo de Markov escondido

2.1.3 Predicción, filtrado y suavizado

Las técnicas de asimilación de datos buscan hacer inferencia estadística en state-space models, es decir que la distribución de interés es $p(\mathbf{x}|\mathbf{y})$. Sin embargo, dado que tenemos muchas realizaciones en el tiempo para $\mathbf{x}$ e $\mathbf{y}$, debemos ser más específicos. Habitualmente distinguimos 3 distribuciones objetivo de interés:

- La distribución predictiva (también llamada de pronóstico o forecast) $p(\mathbf{x}_t|\mathbf{y}_{1:s})$ con $s < t$. Esta es la distribución de un estado "futuro" usando datos del "pasado"
- La distribución filtrante (también llamada análisis) $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ que informa sobre el estado actual usando observaciones pasadas y actuales
- La distribución suavizante $p(\mathbf{x}_t|\mathbf{y}_{1:s})$ con $s > t$ que puede ser interpretada como un reanálisis del estado habiendo colectado observaciones futuras al momento sobre el que se hace inferencia.

2.1.4 Algoritmo *forward-backward*

En modelos de Markov escondidos, bajo la suposición de que contamos con un modelo de la distribución inicial del estado $p(\mathbf{x}_0)$, el modelo de transición $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ y el modelo observacional $p(\mathbf{y}_t|\mathbf{x}_t)$ se puede deducir un algoritmo para obtener de manera secuencial las distribuciones suavizantes. Además, como un subproducto se obtienen las distribuciones filtrantes y las de pronóstico con un grado de separación temporal.

Si consideramos una ventana de tiempo $t = 1, ..., T$ el algoritmo primero realiza el forward-pass alternando un paso de predicción, en el que obtiene $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$ con un paso de filtrado (también llamado análisis o update) en el que se incorpora la información de la observación a tiempo t y se obtiene $p(\mathbf{x}_t|\mathbf{y}_{1:t})$.

Para $t = 1, ..., T$:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})d\mathbf{x}_{t-1} \qquad \text{Predicción} \qquad (2.7)$$

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) \qquad \text{Análisis} \qquad (2.8)$$

Para hacer la predicción se integra utilizando el modelo de transición. La interpretación de la fórmula es que se calcula la probabilidad de $\mathbf{x}_t$ dado $\mathbf{x}_{t-1}$ considerando todos los valores posibles de $\mathbf{x}_{t-1}$ que obedecen a la distribución filtrante del tiempo anterior. De esta manera, el paso de predicción es el encargado de propagar hacia adelante la distribución del estado. Por otro lado, para hacer el análisis se usa la

regla de Bayes y se incorpora $\mathbf{y}_t$ utilizando el modelo observacional, es decir se actualiza la distribución obtenida en el paso de predicción. Notemos que usamos la convención notacional $\mathbf{y}_{1:0} = \emptyset$ lo que le da consistencia a las fórmulas para $t = 0$.

Las distribuciones obtenidas en el forward-pass pueden ser utilizadas a su vez para computar las distribuciones suavizantes iterando esta vez hacia atrás, desde el último tiempo hacia el primero de la siguiente manera:

Para $t = T - 1, ..., 0$:

$$p(\mathbf{x}_t|\mathbf{y}_{1:T}) = p(\mathbf{x}_t|\mathbf{y}_{1:t}) \int \frac{p(\mathbf{x}_{t+1}|\mathbf{x}_t)}{p(\mathbf{x}_{t+1}|\mathbf{y}_{1:t})} p(\mathbf{x}_{t+1}|\mathbf{y}_{1:T}) d\mathbf{x}_{t+1} \qquad \text{Suavizado} \qquad (2.9)$$

donde el caso $t = T$ ya está cubierto pues la distribución filtrante para el último tiempo coincide con la suavizante.

La deducción de las fórmulas para predicción, análisis y suavizado están desarrolladas en mayor detalle en el apéndice A.1.

A pesar de dar una forma general de resolver el problema que plantea la asimilación de datos, en la práctica su aplicación no es tan directa. La integración sobre el espacio de las variables de estado es en general privativa incluso en espacios de dimensionalidad mediana. Por otro lado, hemos hecho la suposición de que contamos con la probabilidad de transición $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ y esto usualmente no es el caso. El modelo de transición $\mathcal{M}_t$ suele funcionar como una caja negra, de manera que contamos con una forma de muestrear $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ pero no necesariamente de evaluar la función de densidad de probabilidad para calcular las integrales necesarias. Existe una gran diversidad de técnicas de asimilación de datos que abordan este problema de distintas maneras. En las secciones subsiguientes describiremos las más relevantes.

2.2 Filtro de Kalman

El filtro de Kalman trabaja sobre una simplificación del problema dado por las Ecuaciones 2.1 y 2.2. Se asume que el modelo de transición de las variables de estado y el modelo observacional son lineales y que las componentes estocásticas se manifiestan como errores Gaussianos aditivos insesgados. Esto resulta en la siguiente reformulación de las ecuaciones:

$$\mathbf{x}_t = \mathbf{M}_t\mathbf{x}_{t-1} + \boldsymbol{\eta}_t, \qquad (2.10)$$

$$\mathbf{y}_t = \mathbf{H}_t\mathbf{x}_t + \boldsymbol{\nu}_t. \qquad (2.11)$$

donde $\mathbf{M}_t$ y $\mathbf{H}_t$ son operadores lineales y $\boldsymbol{\eta}_t$ y $\boldsymbol{\nu}_t$ son variables aleatorias Gaussianas con media $\mathbf{0}$ y matrices de covarianza $\mathbf{Q}_t$ y $\mathbf{R}_t$ respectivamente, es decir $\boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_t)$ y $\boldsymbol{\nu}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_t)$. Esta configuración del problema asume que tanto el error de modelo como el observacional son insesgados y quedan codificados por completo en las matrices $\mathbf{Q}_t$ y $\mathbf{R}_t$.

Si además suponemos que la distribución inicial de $\mathbf{x}_0$ es Gaussiana, entonces las distribuciones predictivas y filtrantes serán también Gaussianas. Esto es porque en el paso de predicción, la linealidad del operador de transición preserva la Gaussianidad, lo cual resulta en que en la aplicación de la regla de Bayes en el paso de análisis tengamos verosimilitud y *prior* Gaussianas resultando en una distribución *a posteriori* (la filtrante) también Gaussiana. Este tipo de distribución tiene la propiedad de que pueden ser representadas de manera completa a través de dos parámetros: su vector de medias y su matriz de covarianza. Por lo tanto, la tarea del filtro de Kalman es producir secuencias de medias y covarianzas predictivas, $\{\mathbf{x}_t^f, \mathbf{P}_t^f\}_{t=1}^T$ y medias y

covarianzas filtrantes $\{\mathbf{x}_t^a, \mathbf{P}_t^a\}_{t=1}^T$, de manera que:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) \sim \mathcal{N}(\mathbf{x}_t^f, \mathbf{P}_t^f)$$
$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \sim \mathcal{N}(\mathbf{x}_t^a, \mathbf{P}_t^a)$$

Si incorporamos las densidades de probabilidad Gaussianas en las fórmulas de predicción 2.7 y análisis 2.8 del *forward-pass* se obtienen ecuaciones cerradas para la secuencia de medias y matrices de covarianza de las distribuciones predictivas y filtrantes. Las ecuaciones que se obtienen para el pronóstico son:

$$\mathbf{x}_t^f = \mathbf{M}_t \mathbf{x}_{t-1}^a \tag{2.12}$$
$$\mathbf{P}_t^f = \mathbf{Q}_t + \mathbf{M}_t \mathbf{P}_{t-1}^a \mathbf{M}_t^T \tag{2.13}$$

mientras que para el análisis resulta:

$$\mathbf{x}_t^a = \mathbf{x}_t^f + \mathbf{K}_t(\mathbf{y}_t - \mathbf{H}_t \mathbf{x}_t^f) \tag{2.14}$$
$$\mathbf{P}_t^a = (\mathbf{I} - \mathbf{K}_t \mathbf{H}_t)\mathbf{P}_t^f \tag{2.15}$$

donde $\mathbf{K}_t = \mathbf{P}_t^f \mathbf{H}_t^T (\mathbf{R}_t + \mathbf{H}_t \mathbf{P}_t^f \mathbf{H}_t^T)^{-1}$ se denomina matriz de ganancia de Kalman, mientras a las diferencias $(\mathbf{y}_t - \mathbf{H}_t \mathbf{x}_t^f)$ se las denominada innovaciones porque dan cuenta de la diferencia entre el pronóstico y la observación. La deducción de estas fórmulas está desarrollada en el apéndice A.2

Notemos que la media de los pronósticos es tan solo la propagación hacia adelante de la media filtrante del tiempo anterior. Por otro lado la matriz de ganancia de Kalman funciona como una matriz de pesos que determina si el estado del análisis será más cercano al pronóstico o si le dará más importancia a la observación. Para dar una interpretación de esto consideremos un ejemplo sencillo en donde el espacio observacional y de las variables de estado son unidimensionales. Bajo esta suposición las matrices $\mathbf{P}^f$ y $\mathbf{R}$ se reducen a escalares: la varianza del error del pronóstico, σ_f^2, y del error observacional σ_{obs}^2. Además consideraremos que $\mathbf{H}$ es la identidad, es decir que es igual al escalar 1. Notemos que quitamos la dependencia temporal, sólo consideraremos el paso de un pronóstico a un análisis. De esta manera las ecuaciones que determinan a la ganancia de Kalman y a la media y varianza del análisis se reducen a las siguientes expresiones:

$$K = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_{obs}^2} \tag{2.16}$$
$$x^a = x^f + K(y - x^f) \tag{2.17}$$
$$P^a = (1 - K)P^f \tag{2.18}$$

Podemos ver que si el error en el pronóstico es muy pequeño en relación al error observacional, $\sigma_f^2 \ll \sigma_{obs}^2$, entonces K tiende a 0 con lo que x^a tiende a x^f y P^a a P^f. Es decir que si el pronóstico es mucho más preciso en relación a la observación, el análisis va a ignorar la información de la observación y será prácticamente igual al pronóstico. De manera análoga, si el error observacional es mucho menor en relación al del pronóstico, es decir si $\sigma_{obs}^2 \ll \sigma_f^2$, K tiende a la identidad y por lo tanto x^a tiende a y mientras que P^a tiende a 0. Esto significa que si la observación es muy precisa respecto al pronóstico el análisis tenderá a la observación y con varianza nula pues y es una realización puntual.

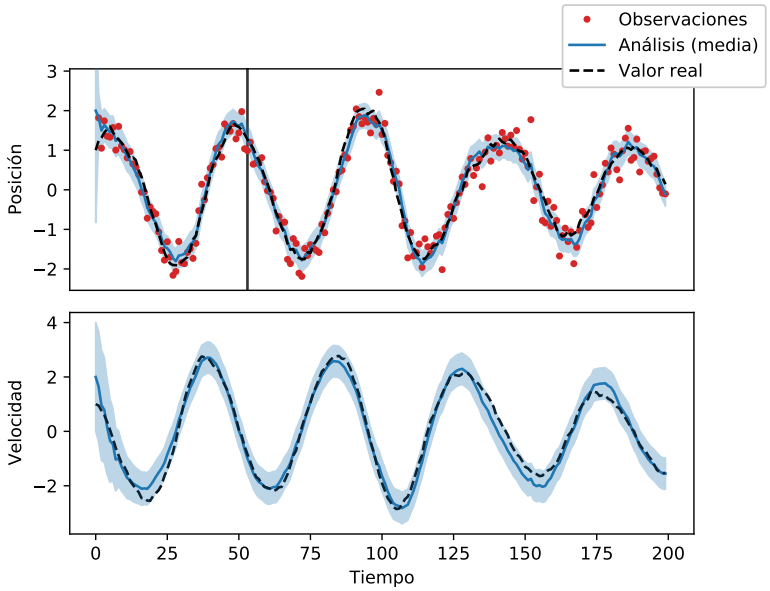

Figura 2.3: Trayectorias reales y estimadas mediante el filtro de Kalman. Sólo la posición es observada

Ejemplo: oscilador armónico

Consideramos aquí un sencillo modelo de un oscilador armónico y ejemplificamos cómo usar el filtro de Kalman. Estos sistemas se pueden modelar con la siguiente ecuación diferencial:

$$\frac{\partial^2 x}{\partial t} = -\omega^2 x(t)$$

donde ω es la frecuencia angular. Podemos considerar un integrador numérico de Euler semi-implícito para la posición x y la velocidad v en tiempo discretizado a intervalos de longitud dt, con lo que obtenemos:

$$x_t = (1 - \omega^2 dt^2)x_{t-1} + v_{t-1}dt$$
$$v_t = v_{t-1} - \omega^2 x_{t-1}dt$$

Esto constituye un sistema lineal que admite la expresión:

$$\underbrace{\begin{pmatrix} x_t \\ v_t \end{pmatrix}}_{\mathbf{x}_t} = \underbrace{\begin{pmatrix} 1-\omega^2 dt^2 & dt \\ -\omega^2 dt & 1 \end{pmatrix}}_{\mathbf{M}_t} \cdot \underbrace{\begin{pmatrix} x_{t-1} \\ v_{t-1} \end{pmatrix}}_{\mathbf{x}_{t-1}} \tag{2.19}$$

Adicionalmente incorporamos al modelo ruido Gaussiano aditivo constante $\boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$ para obtener una expresión como 2.10. Para el modelo observacional consideraremos que sólo obtenemos datos de la posición utilizando la matriz $\mathbf{H}_t = (1, 0)$ y que estos datos tienen error aditivo $\boldsymbol{\nu}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$.

Simulamos entonces una trayectoria evolucionando temporalmente la Ecuación 2.19 desde una condición inicial arbitraria. Luego, añadimos ruido a esta trayectoria para representar el error observacional, y obtenemos observaciones sintéticas. Utilizando el filtro de Kalman podemos intentar estimar la trayectoria real a través de la asimilación de los datos simulados. En la Figura 2.3 podemos ver las trayectorias reales (que no tienen amplitud constante debido al error introducido por $\boldsymbol{\eta}_t$) y

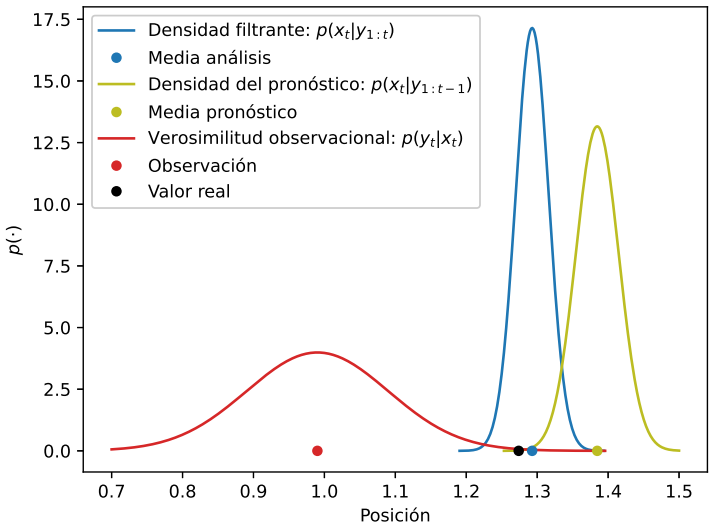

Figura 2.4: Funciones de densidad de probabilidad filtrante y de pronóstico junto a la verosimilitud de la observación y el valor real. Corresponde al tiempo indicado por el corte vertical en la Figura 2.3

las estimaciones de las medias del filtro. Estas son naturalmente más precisas para la variable observada, sin embargo las correlaciones entre las observaciones y las variables no observadas que utiliza el filtro permiten un seguimiento aproximado de la velocidad. Es importante señalar que el filtro de Kalman no produce solo una estimación de las medias sino que provee una medida de la incerteza mediante estimaciones de las varianzas. En la Figura 2.3 graficamos los intervalos de confianza del 95% determinados por la estimación de la varianza filtrante determinada por P_t^a. Esta se reduce en los primeros ciclos de asimilación. El sistema va incorporando las observaciones una a una de manera secuencial para obtener las estimaciones de las distribuciones filtrantes, y a mediada que lo hace la varianza del error se reduce. Las distribuciones filtrantes $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ están condicionadas a más observaciones a medida que t crece y como consecuencia en las primeras iteraciones tenemos mayor varianza hasta que se acumulan suficientes observaciones y la variabilidad se estabiliza. En la Figura 2.4 se representan las densidades de probabilidad del pronóstico y del análisis junto con la verosimilitud de la observación proyectadas sobre el espacio de x. Corresponden a un tiempo fijo indicado por la línea de corte vertical del panel superior de la Figura 2.3. El pronóstico actúa como una probabilidad *a priori* que se combina con la verosimilitud, mediante la regla de Bayes que subyace al análisis del filtro de Kalman, para conformar la distribución *a posteriori* que corresponde al análisis. La estimación del análisis mejora al pronóstico utilizando la observación. Además, la incerteza del análisis es menor a la del pronóstico y a la de la observación, lo cual es una propiedad que se hereda de la regla de Bayes.

2.3 Métodos variacionales

El problema de obtener un análisis a partir de un pronóstico y una observación puede ser interpretado de manera variacional. A pesar de que no nos centramos en este tipo de metodología en este trabajo, mencionamos brevemente la idea que motiva a los métodos variacionales pues constituyen una familia de técnicas que

han sido aplicadas a modelos meteorológicos operacionales y que además dan una perspectiva que enriquece la interpretación de otras estrategias de asimilación de datos. Para esto consideremos el problema de obtener un análisis partiendo desde un pronóstico de media $\mathbf{x}^f$ y un error de varianza $\mathbf{P}^f$, una observación $\mathbf{y}$ con un error de varianza $\mathbf{R}$ y con el operador $\mathcal{H}$ que mapea el espacio de las variables de estado al espacio de las observaciones. Nuestro objetivo será encontrar el estimador MAP (máximo *a posteriori*), es decir tomaremos a la media del análisis, $\mathbf{x}^a$ como el valor de $\mathbf{x}$ tal que maximice la probabilidad a posteriori $p(\mathbf{x}|\mathbf{y})$. Aplicando la regla de Bayes tenemos que:

$$\begin{aligned}\mathbf{x}^a &= \underset{\mathbf{x}}{\operatorname{argmax}} \{p(\mathbf{x}|\mathbf{y})\} \\ &= \underset{\mathbf{x}}{\operatorname{argmax}} \left\{ \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} \right\} \\ &= \underset{\mathbf{x}}{\operatorname{argmin}} \{-\log p(\mathbf{y}|\mathbf{x}) - \log p(\mathbf{x})\}\end{aligned}$$

Si adicionalmente suponemos que la verosimilitud $p(\mathbf{y}|\mathbf{x})$ y el *prior* $p(\mathbf{x})$ son Gaussianas podemos reescribir el problema como la minimización de una función de costo:

$$\mathbf{x}^a = \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ (\mathbf{x} - \mathbf{x}^f){\mathbf{P}^f}^{-1}(\mathbf{x} - \mathbf{x}^f)^T + (\mathbf{y} - \mathcal{H}(\mathbf{x}))\mathbf{R}^{-1}(\mathbf{y} - \mathcal{H}(\mathbf{x}))^T \right\} \tag{2.20}$$

$$= \underset{\mathbf{x}}{\operatorname{argmin}} \{J(\mathbf{x})\} \tag{2.21}$$

El primer término da cuenta del costo cuadrático asociado a alejarse del pronóstico y, análogamente, el segundo corresponde a la penalización por alejarse de la observación. Es de esperar entonces que el minimizador de J pondere ambas fuentes de información. De hecho, mientras menor sea el error del pronóstico, mayor será el costo de alejarse de este y para la observación tenemos una situación equivalente. Notemos que esta función de costo supone que los errores sean Gaussianos, aditivos e insesgados. Por otro lado, no hay suposiciones explícitas para el modelo observacional $\mathcal{H}$ ni para el modelo $\mathcal{M}$. Este último, ni siquiera está incluido porque se hace la suposición de que ya se cuenta con un pronóstico. Sin embargo, si el pronóstico proviene de un modelo no lineal, no es posible en principio garantizar la Gaussianidad de su error. Sobre el operador observacional también hay que aclarar que dependiendo del método con que se minimice la función de costo, es posible que se necesiten requisitos adicionales: por ejemplo sería esperable que el método de optimización requiera que el operador sea linealizable para poder utilizar métodos de segundo orden. Bajo la hipótesis que el operador observacional es lineal, J resulta en una función cuadrática con un único mínimo el cual coincide con la media del análisis tal como se lo computa en el filtro de Kalman. Por otro lado, cuando $\mathcal{H}$ es no lineal la función de costo captura estos efectos y de hecho puede constituir una función con múltiples mínimos locales que no sea tan sencilla para optimizar.

La metodología que implementa la minimización de J es comúnmente conocida como 3DVAR (Courtier et al., 1998), pero hay que mencionar que el planteo variacional del problema abre la puerta a múltiples variantes, ya que se pueden explorar distintos tipos de optimizadores (por ejemplo simplex, quasi-Newton o gradiente conjugado), distintas condiciones iniciales, precondicionamientos, regularización de J, etc. Finalmente mencionamos el método 4DVAR (Rabier y Liu, 2003; Talagrand y Courtier, 1987) el cual es una extensión de 3DVAR pero que procesa una ventana de observaciones simultáneamente, es decir funciona como un suavizador en lugar de

como un filtro. El problema que busca optimizar 4DVAR es la minimización de una función de costo similar a 2.20 pero que incluye muchas observaciones distribuidas temporalmente. Además se plantea como un problema de optimización con restricciones pues la minimización se hace sujeta a que la solución sea una trayectoria del modelo del que se obtienen los pronósticos.

2.4 Técnicas por ensambles

El filtro de Kalman constituye una técnica robusta que da una solución exacta en el caso de modelos lineales con errores Gaussianos aditivos. En ciertos casos es posible considerar linealizaciones de los operadores $\mathcal{M}_t$ y $\mathcal{H}_t$ y aplicar el filtro de Kalman tradicional con estas aproximaciones. Este método se denomina filtro de Kalman extendido y también producirá estimaciones de las medias y covarianzas predictivas y filtrantes. Aún así, estas dos técnicas no dan respuesta a dos situaciones frecuentes en las aplicaciones de asimilación de datos. Por un lado, es factible que el modelo no sea linealizable, ya sea porque es tratado como una caja negra o porque la aproximación lineal es imprecisa. En estas situaciones, los pronósticos serán no Gaussianos y es necesario utilizar técnicas que permitan representar otro tipo de distribuciones. Por otro lado, en modelos meteorológicos es común que el espacio de las observaciones tenga alta dimensionalidad ($\sim 10^5$) y el de las variables de estado aún más ($\sim 10^7$) por lo que computar y almacenar las matrices de covarianza $\mathbf{P}_t^f$ y $\mathbf{P}_t^a$ puede ser prohibitivo (Katzfuss et al., 2016). Para dar cuenta de estos problemas se pueden usar técnicas basadas en partículas o ensambles. Estos, en lugar de representar las distribuciones objetivo a través de sus parámetros como en el filtro de Kalman y el filtro de Kalman extendido, se busca representarlas a través de muestras, es decir un ensamble de puntos en el espacio de las variables de estado. Cada punto muestral suele ser denominado partícula o miembro de ensamble de acuerdo a la técnica en cuestión.

Vamos a introducir aquí dos familias de métodos basados en ensambles: los filtros de partículas y los filtros de Kalman por ensambles (EnKFs). Los filtros de partículas permiten, en principio, la representación de distribuciones no paramétricas con formas arbitrarias por lo que pueden ser utilizados en escenarios no Gaussianos. Por otro lado, los EnKFs son habitualmente utilizados para mapear el problema al espacio que generan los miembros del ensamble, el cual tiene una dimensionalidad en general mucho menor que el de las variables de estado haciendo posible el cómputo. Es importante aclarar que ninguna de estas técnicas provee una solución *off-the-shelf* para problemas de asimilación de datos arbitrarios e incluso cada método trae aparejado un conjunto de dificultades técnicas (por ejemplo la degeneración de pesos en el filtro de partículas o el colapso del ensamble en EnKFs). Tanto para filtros de partículas como EnKFs existe una amplia gama de variaciones e implementaciones que introducen características particulares o que buscan resolver o mitigar algún problema en particular. Comenzaremos introduciendo los filtros de partículas porque estos dan una noción clara de por qué tiene sentido usar muestras para representar distribuciones para luego seguir con los EnKFs.

2.4.1 Monte Carlo secuencial

Los filtros de partículas son también conocidos como métodos de Monte Carlo secuencial ya que utilizan el esquema secuencial de dos pasos de predicción-análisis descrito por las Ecuaciones 2.7 y 2.8. De hecho, el enfoque de estos métodos es el

de resolver las integrales de estas ecuaciones, no de manera explícita sino mediante aproximaciones de Monte Carlo. Esto significa que si tenemos una función de densidad de probabilidades p y un conjunto de partículas $\{\mathbf{x}^{(i)}\}_{i=1}^{N}$ muestreadas de manera independiente con esta probabilidad, i.e. $\mathbf{x}^{(i)} \sim p$, entonces, por la ley de los grandes números, se pueden aproximar valores esperados en base a la distribución p utilizando la media muestral de las partículas:

$$\frac{1}{N}\sum_{i=1}^{N} f(\mathbf{x}^{(i)}) \xrightarrow{c.s.} \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

Otra interpretación de la aproximación de Monte Carlo es que se utiliza la aproximación empírica de p basada en la muestra $\{\mathbf{x}^{(i)}\}_{i=1}^{N}$, es decir que se considera lo siguiente:

$$p(\mathbf{x}) \approx \frac{1}{N}\sum_{i=1}^{N} \delta_{\mathbf{x}^{(i)}}(\mathbf{x})$$

donde $\delta_{\mathbf{a}}$ es la delta de Dirac que acumula toda la probabilidad en el punto $\mathbf{a}$ siendo nula en todo otro punto (ver por ejemplo Doucet et al., 2001).

Existe una generalización de la aproximación de Monte Carlo denominada muestreo de importancia. A pesar de su nombre es un método para aproximar integrales. Cuando no es posible muestrear p se puede entonces muestrear otra distribución con densidad q que actúe de proxy de p. En principio la única condición para q es que su soporte contenga al de p. Este método también admite que no se pueda evaluar exactamente p y q sino que sólo podamos evaluar versiones no normalizadas $\tilde{p}$ y $\tilde{q}$ de estas densidades. Entonces, si tenemos un conjunto de partículas $\{\mathbf{x}^{(i)}\}_{i=1}^{N}$ tales que $\mathbf{x}^{(i)} \sim q$, podemos hacer la aproximación dada por las siguientes fórmulas:

$$\begin{aligned}\int f(\mathbf{x})p(\mathbf{x})d\mathbf{x} &\approx \sum_{i=1}^{N} w_i f(\mathbf{x}^{(i)}) \\ w_i &= \tilde{w}_i / \textstyle\sum_{i=1}^{N} \tilde{w}_i \\ \tilde{w}_i &= \tilde{p}(\mathbf{x}^{(i)})/\tilde{q}(\mathbf{x}^{(i)})\end{aligned}$$

donde w_i son denominados pesos de importancia y $\tilde{w}_i$ son sus versiones no normalizadas. Por su parte, q es denominada la distribución *propuesta*. La metodología puede incluso resultar más eficiente en el número de partículas necesarias que el método de Monte Carlo tradicional puesto que se puede elegir una distribución q que muestree mejor el espacio en relación a la función f. En general es conveniente muestrear más partículas en las regiones en las que el integrando sea más relevante, es decir donde $p(\mathbf{x})q(\mathbf{x})$ sea mayor (ver por ejemplo MacKay, 2003; Murphy, 2012).

Los métodos de Monte Carlo secuencial producen conjuntos de partículas que aproximan a la distribución filtrante. Estas partículas además pueden ser ponderadas, es decir pueden traer pesos asociados. Concretamente, para cada tiempo t se obtienen $\{\mathbf{x}_t^{(i)}, w_t^{(i)}\}_{i=1}^{N}$ tal que sea una aproximación empírica de $p(\mathbf{x}_t|\mathbf{y}_{1:t})$.

Los filtros de partículas obtienen estas muestras en dos pasos básicos: primero se muestrean partículas de una distribución propuesta q y luego se computan sus pesos. Con una elección adecuada de q las partículas al tiempo t pueden ser obtenidas a partir de las partículas a tiempo $t-1$ y los pesos pueden ser computados como una actualización de los pesos del tiempo anterior. En particular, dada una muestra pesada $\{\mathbf{x}_{t-1}^{(i)}, w_{t-1}^{(i)}\}_{i=1}^{N}$ correspondiente al tiempo $t-1$, el procedimiento consiste

en:

- Muestrear partículas:

$$\mathbf{x}_t^{(i)} \sim q(\mathbf{x}_t|\mathbf{x}_{t-1}^{(i)}, \mathbf{y}_t)$$

- Actualizar pesos:

$$w_t^{(i)} \propto w_{t-1}^{(i)} \frac{p(\mathbf{y}_t|\mathbf{x}_t^{(i)})p(\mathbf{x}_t^{(i)}|\mathbf{x}_{t-1}^{(i)})}{q(\mathbf{x}_t^{(i)}|\mathbf{x}_{t-1}^{(i)}, \mathbf{y}_t)}$$

Esta implementación es habitualmente llamada SIS (*sequential importance sampling*) y en la práctica tiene el problema de que los pesos suelen concentrarse en una sola partícula, es decir uno de los pesos es prácticamente 1 mientras que el resto es prácticamente 0. Este efecto se denomina degeneración del filtro y es indeseable pues la representación muestral de la distribución filtrante pierde expresividad. Además de aumentar el número de partículas existen varias estrategias para mitigar este efecto, la más conocida de ellas es el remuestreo. Este método consiste en muestrear con reemplazo las partículas de la distribución empírica dada por los pesos. Es decir, una vez computados los pesos se obtiene un nuevo conjunto de partículas con pesos uniformes:

$$\hat{\mathbf{x}}_t^{(i)} \sim \sum_{i=1}^{N} w_i \delta_{\mathbf{x}^{(i)}}$$
$$\hat{w}_i = 1/N$$

Esto tiene el efecto que las partículas con mayor peso estarán repetidas y las de menos peso serán eliminadas. A pesar de mitigar la degeneración del filtro causa el problema de empobrecimiento de diversidad del que hablaremos más adelante. No es necesario hacer remuestreo de partículas en cada paso de tiempo y un criterio común para decidir si se hace o no es verificar si el número efectivo de partículas N_{eff} está por debajo de un valor umbral N_T. El número efectivo de partículas se puede estimar de la siguiente manera (Liu y Chen, 1998):

$$N_{eff} \approx \frac{1}{\sum_{i=1}^{N}(w_i)^2} \doteq \widehat{N_{eff}}$$

Este filtro de partículas que incorpora remuestreo suele ser denominado SIR (*sequential importance resampling*) y lo expresamos en forma algorítmica en el Algoritmo 1.

Una de las implementaciones más sencillas de este tipo de filtro es el llamado *bootstrap* (Gordon et al., 1993) y consiste en tomar la distribución propuesta como la probabilidad de transición, i.e., $q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_t) = p(\mathbf{x}_t|\mathbf{x}_{t-1})$. Esto significa que el muestreo de partículas es simplemente aplicar el modelo de transición a todas las partículas del tiempo anterior. Además, las implementaciones más comunes del filtro *bootstrap* hacen remuestreo en cada paso de tiempo. Esto significa que los pesos son sólo calculados para hacer remuestreo pero las partículas que representan a la distribución filtrante tienen pesos uniformes y no hay que almacenarlos. Además esto simplifica el cómputo de los pesos a la expresión $w_t^{(i)} \propto p(\mathbf{y}_t|\mathbf{x}_t^{(i)})$. En el Algoritmo 2 especificamos este método.

La elección de la distribución de transición como distribución propuesta en el filtro de partículas *bootstrap* significa que el muestreo de partículas no es más que

Algoritmo 1: Filtro de partículas SIR

Muestrear partículas iniciales: $\{\mathbf{x}_0^{(i)}\}_{i=1}^{N_p} \sim p(\mathbf{x}_0)$
for $t = 1, ..., T$ **do**
  Muestrear partículas de la distribución propuesta:
    $\widehat{\mathbf{x}}_t^{(i)} \sim q(\mathbf{x}_t|\mathbf{x}_{t-1}^{(i)}, \mathbf{y}_t)$
  Actualizar y normalizar pesos:
    $w_t^{(i)} \propto w_{t-1}^{(i)} \frac{p(\mathbf{y}_t|\widehat{\mathbf{x}}_t^{(i)})p(\widehat{\mathbf{x}}_t^{(i)}|\mathbf{x}_{t-1}^{(i)})}{q(\widehat{\mathbf{x}}_t^{(i)}|\mathbf{x}_{t-1}^{(i)}, \mathbf{y}_t)}$
  Aproximar el número efectivo de partículas:
    $\widehat{N_{eff}} = \frac{1}{\sum_{i=1}^{N_p}(w_i)^2}$
  if $\widehat{N_{eff}} < N_T$ **then**
    $\mathbf{x}_t^{(i)} \sim \sum_{i=1}^{N_p} w_i \delta_{\widehat{\mathbf{x}}_t^{(i)}}$
    $w_t^{(i)} = 1/N_p$
  end
end

evolucionar cada partícula del tiempo anterior utilizando el modelo. Las partículas muestreadas constituyen entonces un pronóstico. Esta interpretación no es generalizable al filtro SIR general. Para indicar esto en el Algoritmo 2 usamos el supraíndice f (por *forecast*) en las partículas del pronóstico y a en las de análisis.

El cómputo de los pesos está dado por la verosimilitud observacional por lo que estos cuantifican cuan afín es cada partícula a la observación. Por su parte, la forma de transformar al conjunto de partículas de pronóstico en una muestra filtrante es a través del remuestreo. El remuestreo, al ser con reemplazo, tiene el efecto de multiplicar las partículas con pesos altos (cercanas a la observación) y eliminar las partículas con pesos bajos (lejanas a la observación). Este mecanismo se suele comparar con la selección natural, sólo sobreviven y se reproducen las partículas "más aptas".

Como se anticipó, el remuestreo introduce otro problema: el empobrecimiento de diversidad. Esto se debe a que tendremos muchas réplicas de las partículas con más peso, es decir una muestra que cubre pobremente el espacio muestral. Este efecto es claramente indeseable y para paliarlo es habitual introducir, luego del remuestreo, un paso en el que las partículas se mutan o se mueven de alguna forma para no tener tantas partículas repetidas. El término mutación proviene de la interpretación del filtro de partículas *bootstrap* como un mecanismo de selección del más apto. También se suele denominar *jittering*. Esta modificación de las partículas se puede hacer incorporando pasos de MCMC. También es posible mitigar el problema mediante regularización, utilizando *kernel density estimation* sobre las distribuciones empíricas (Arulampalam et al., 2002; Ruchi et al., 2019; Särkkä, 2013).

La diversidad de filtros de partículas es vasta: la transformación de un conjunto de partículas que representa un pronóstico en una muestra de la probabilidad *a posteriori* puede ser lograda con diversos enfoques. Muchos de ellos buscan mitigar los problemas típicos de los filtros de partículas más tradicionales. Por ejemplo, en Reich, 2013 se utiliza transporte óptimo en distribuciones discretas de manera que se asegura que las partículas resultantes tengan todas los mismos pesos. Otra alternativa para evitar que los pesos sean muy pequeños es lo que se llama el temperado

Algoritmo 2: Filtro de partículas bootstrap

Muestrear partículas iniciales: $\{\mathbf{x}_0^{a,(i)}\}_{i=1}^{N_p} \sim p(\mathbf{x}_0)$
for $t = 1, ..., T$ **do**
  Evolucionar partículas:
    $\mathbf{x}_t^{f,(i)} \sim p(\mathbf{x}_t|\mathbf{x}_{t-1}^{a,(i)})$ para $i = 1, ..., N_p$
  Actualizar y normalizar pesos:
    $w^{(i)} \propto p(\mathbf{y}_t|\mathbf{x}_t^{f,(i)})$
  Remuestrear:
    $\mathbf{x}_t^{a,(i)} \sim \sum_{i=1}^{N_p} w_i \delta_{\mathbf{x}^{f,(i)}}$
end

de la verosimilitud (Neal, 1996). En general esto se traduce en que en lugar de una transformación de un paso del pronóstico al análisis, se hace una transformación más suave en varios pasos comenzando con una verosimilitud más ancha y progresivamente acercarse a la verosimilitud real. Esto hace que las partículas se acercan de a poco a las zonas de mayor verosimilitud. Existen también filtros que combinan el paso de análisis de un filtro de Kalman por ensambles con el de un filtro de partículas tradicional (Frei y Künsch, 2013; Stordal et al., 2011). Estos enfoques están fuertemente relacionados a interpretaciones de las distribuciones como mixturas de Gaussianas. Como mencionamos, el temperado considera una transición de pronóstico a análisis en muchos pasos; si generalizamos este concepto a infinitos pasos se puede comenzar a pensar en flujos de partículas, es decir filtros que mueven dinámicamente las partículas hacia la distribución *a posteriori* de acuerdo a un flujo y de acuerdo al flujo de que se elija para el movimiento de las partículas tendremos diferentes implementaciones. Una opción es tomar actualizaciones de acuerdo a la dinámica de Langevin (Liu, 2017) y otra fuertemente emparentada es utilizar decenso de gradiente variacional de Stein Liu y Wang, 2016 con lo cual se puede derivar el filtro de partículas de mapeo variacional (Pulido y van Leeuwen, 2019) del que hablaremos más adelante. En Van Leeuwen et al., 2019 se puede encontrar un buen reporte sobre las diversas implementaciones de filtros de partículas.

El filtro de partículas SIR tiene la ventaja de hacer pocos requisitos sobre el modelo. En el caso particular del filtro *bootstrap* sólo se necesita evaluar la verosimilitud observacional y muestrear de la probabilidad de transición. Obtener muestra de la probabilidad de transición no es otra cosa que evolucionar el modelo, y este requisito por sí solo significa que el modelo puede ser tratado como una caja negra, una característica deseable en muchas situaciones. Como no hay ninguna suposición sobre las distribuciones, estos métodos pueden ser utilizados en escenarios de no-Gaussianidad y no linealidad. Por otro lado, este tipo de filtros habitualmente necesitan muchas partículas para tener una buena performance y además padecen de los problemas ya mencionados de degeneración del filtro y empobrecimiento de la diversidad. También es común que muchos pesos sean muy pequeños, lo cual puede causar problemas numéricos. Estos problemas se acentúan en espacios de alta dimensionalidad por lo que en estas situaciones es habitual utilizar filtros de Kalman por ensambles, los cuales introducimos más adelante. Los filtros de partículas en alta dimensionalidad son un desafío actualmente pero existen importantes avances por ejemplo aplicando localizacion (Vossepoel y Jan van Leeuwen, 2007), una herramienta normal en filtros de Kalman por ensambles. También existe la posibilidad de mover inteligentemente las partículas a zonas donde los pesos no se degeneren lo cual puede lograrse mediante filtros por flujos de partículas.

VMPF

En el trabajo de investigación de esta tesis utilizamos un filtro basado en flujo de partículas llamado filtro de partículas de mapeo variacional (VMPF, Pulido y van Leeuwen, 2019), para utilizar en combinación con las técnicas de tratamiento de errores observacionales y de modelo desarrolladas. Hacemos una breve mención sobre la idea que subyace a este método. En los filtros basados en flujo de partículas se considera que estas obedecen en un pseudotiempo s a una ecuación diferencial

$$\frac{d\mathbf{x}}{ds} = f_s(\mathbf{x})$$

de manera que las partículas en el pseudotiempo inicial se distribuyen como la *prior* y en el pseudotiempo final obedecen a la distribución *a posteriori* pasando por distribuciones intermedias p_s. En particular, el VMPF mueve las partículas en la dirección de máximo descenso del gradiente de la divergencia de Kullback-Leibler entre p_s y la distribución objetivo. Además, para evaluar este gradiente se considera una versión kernelizada de la divergencia de Kullback-Leibler. Esto resulta en que, discretizando el pseudotiempo, tenemos:

$$\begin{aligned}\mathbf{x}_{s+1} &= \mathbf{x}_s - \epsilon \nabla KL(\mathbf{x}_s) \\ &= \mathbf{x}_s - \epsilon \int p_s(\mathbf{z})(K(\mathbf{z}, \mathbf{x}_s)\nabla_{\mathbf{z}} \log p(\mathbf{z}|\mathbf{y}) + \nabla_{\mathbf{z}} \mathcal{K}(\mathbf{z}, \mathbf{x}_s))d\mathbf{z}\end{aligned}$$

donde $\mathcal{K}$ es el kernel y la integral que representa al gradiente puede ser aproximada utilizando Monte Carlo basado en la muestra de partículas de la distribución $p_s(\mathbf{z})$ disponible de la iteración correspondiente al pseudotiempo anterior. El parámetro ϵ se conoce como tasa de aprendizaje y existen formas de elegirla de manera dinámica para mejorar la convergencia (Zeiler, 2012). El método resultante tiene la cualidad de que las partículas se mueven hacia la moda de la distribución $p(\mathbf{x}|\mathbf{y})$ pero la amplitud de banda del kernel actúa como una fuerza repulsiva entre las partículas (Liu y Wang, 2016) de manera que la muestra resultante no pierde diversidad. Al utilizar una única partícula, esta se dirige a la moda de la distribución objetivo, dando la estimación MAP (máximo *a posteriori*) la cual, notablemente, es la que se obtiene mediante métodos variacionales como el 3DVAR.

2.4.2 EnKF

Los filtros de Kalman por ensambles utilizan muestras para mantener estimaciones de las distribuciones objetivo pero asumen Gaussianidad de la densidad de predicción y la función de verosimilitud. Bajo esta suposición se incorporan las fórmulas del filtro de Kalman tradicional para actualizar las partículas (en el contexto del EnKF las partículas suelen ser llamadas miembros del ensamble pero utilizaremos las terminologías de manera intercambiable). Notemos que para asumir la Gaussianidad de la densidad de predicción y de la función de verosimilitud es necesario suponer linealidad sobre el modelo de transición y observacional y también errores aditivos Gaussianos: es decir que trabaja sobre el modelo definido por 2.10 y 2.11. En realidad, la suposición de que el modelo de transición es lineal puede ser algo relajada en la práctica pero discutiremos esto más adelante. A pesar de que se pueden obtener buenas estimaciones en el caso de no linealidad moderada en $\mathcal{M}_t$, la técnica pierde robustez si $\mathcal{H}_t$ es no lineal.

Supongamos que a tiempo $t-1$ contamos con un ensamble $\{\mathbf{x}_{t-1}^{a,(i)}\}_{i=1}^{N_p}$ que representa al análisis. Por las suposiciones que hemos hecho esta muestra será Gaussiana y al aplicar el modelo transicional sobre cada uno de estos miembros del ensamble podemos obtener un ensamble $\{\mathbf{x}_t^{f,(i)}\}_{i=1}^{N_p}$ del pronóstico a tiempo t el cual constituirá una muestra Gaussiana. La pregunta entonces, es cómo obtener a partir de este ensamble otro que represente al análisis. La idea del EnKF es aplicar la ecuación de actualización de la media del filtro de Kalman tradicional 2.14 para cada partícula del pronóstico. Esta es la esencia de los EnKF pero para aplicar el concepto correctamente debemos hacer algunas salvedades.

Para aplicar la actualización del filtro de Kalman necesitamos computar $\mathbf{K}_t$ la cual por su parte necesita de la matriz de covarianzas del pronóstico $\mathbf{P}_t^f$. Sin embargo, a diferencia del filtro de Kalman, aquí no propagamos las matrices de covarianzas en el tiempo, sólo una muestra de la distribución y por ello no disponemos de una representación exacta de esta. De cualquier manera, es natural aproximar $\mathbf{P}_t^f$ con la matriz de covarianza muestral del ensamble del pronóstico, $\widehat{\mathbf{P}}_t^f$ para obtener una aproximación $\widehat{\mathbf{K}}_t$ de la matriz de ganancia de Kalman.

Por otro lado, hay que notar que estamos utilizando sobre cada partícula la fórmula de actualización de la media de la distribución, y cada partícula está siendo actualizada en base a la misma observación. Es posible demostrar que si hacemos esto, la covarianza del ensamble del análisis va a estar subestimada. La solución a este problema consiste en actualizar cada partícula usando una perturbación de la observación original. Para obtener el miembro del ensamble $\mathbf{x}_t^{a,(i)}$ utilizaremos la observación modificada $\mathbf{y}_t^{(i)} = \mathbf{y}_t + \mathbf{v}_t^{(i)}$ con $\mathbf{v}_t(i) \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_t)$. El origen de este problema y su solución son planteados en Burgers et al., 1998. El método resultante es denominado EnKF estocástico o EnKF con observaciones perturbadas y en el está descrito Algoritmo 3. En A.4 demostramos que la formulación de este método es correcta, es decir que la media y la covarianza obtenidas son las del análisis.

Algoritmo 3: EnKF estocástico

Muestrear ensamble inicial: $\{\mathbf{x}_0^{a,(i)}\}_{i=1}^{N_p} \sim p(\mathbf{x}_0)$
for $t = 1, ..., T$ **do**
  for $i = 1, ..., N_p$ **do**
    $\boldsymbol{\eta}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_t)$
    $\mathbf{x}_t^{f,(i)} = \mathcal{M}_t(\mathbf{x}_{t-1}^{a,(i)}) + \boldsymbol{\eta}^{(i)}$
  end
  $\widehat{\mathbf{x}}_t^f = \frac{1}{N_p} \sum_{i=1}^{N_p} \mathbf{x}_t^{f,(i)}$
  $\widehat{\mathbf{P}}_t^f = \frac{1}{N_p - 1} \sum_{i=1}^{N_p} (\mathbf{x}_t^{f,(i)} - \widehat{\mathbf{x}}_t^f)(\mathbf{x}_t^{f,(i)} - \widehat{\mathbf{x}}_t^f)^T$
  $\widehat{\mathbf{K}}_t = \widehat{\mathbf{P}}_t^f \mathbf{H}_t^T (\mathbf{R}_t + \mathbf{H}_t \widehat{\mathbf{P}}_t^f \mathbf{H}_t^T)^{-1}$
  for $i = 1, ..., N_p$ **do**
    $\mathbf{v}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_t)$
    $\mathbf{y}_t^{(i)} = \mathbf{y}_t + \mathbf{v}^{(i)}$
    $\mathbf{x}_t^{a,(i)} = \mathbf{x}_t^{f,(i)} + \widehat{\mathbf{K}}_t (\mathbf{y}_t^{(i)} - \mathbf{H}_t \mathbf{x}_t^{f,(i)})$
  end
end

Notemos que la suposición de que el modelo de transición es lineal es la que permite asumir que evolucionar los miembros del ensamble hacia adelante va a conservar la Gaussianidad de la muestra. Sin embargo, a diferencia del filtro de Kalman

tradicional, no precisamos una expresión matricial $\mathbf{M}_t$ del operador. El modelo se usa solamente para propagar las partículas hacia adelante por lo cual podemos relajar la hipótesis de linealidad y pensar en un operador general $\mathcal{M}_t$ que si es levemente no lineal podrá preservar la Gaussianidad en los pronósticos y que además puede ser tratado como una caja negra. De hecho, el EnKF es habitualmente presentado como un método apto para modelos no lineales. El hecho de que no requerimos de ninguna expresión explícita del modelo transicional es una diferencia importante entre el EnKF y el EKF (*extended Kalman filter*). Este útlimo es una extensión del filtro de Kalman tradicional que está diseñado para situaciones de no linealidad y utiliza aproximaciones lineales de los operadores. Esto puede ser un problema puesto que los tangentes lineales de $\mathcal{M}_t$ pueden no estar disponibles.

Además del filtro, existe la posibilidad de obtener representaciones de ensambles de las distribuciones suavizantes utilizando el llamado suavizador de Kalman por ensambles (EnKS). Así como el EnKF se basa en los pasos de predicción y análisis del pase hacia adelante del algoritmo *forward-backward*, la versión del EnKS conocida como Rauch-Tung-Striebel (RTS) consiste en un pase hacia atrás. El Algoritmo 4 muestra el método como es sugerido en Cosme et al., 2012. Notablemente, la metodología se vale solamente de los ensambles de pronóstico y filtrantes producidos por el EnKF por lo que no se necesitan evaluaciones del modelo o del operador observacional. Las partículas suavizantes a tiempo t son computadas como una corrección de las del análisis en las que se incorpora secuencialmente la información de los ensambles al tiempo $t+1$ mediante la ponderación de una matriz de ganancia $\mathbf{K}_t^s$. Esta ganancia se computa utilizando $\mathbf{S}_t^a$ y $\mathbf{S}_{t+1}^f$ las cuales son matrices cuyas columnas son los miembros de los ensambles $\{\mathbf{x}_{t+1}^{f,(i)}\}_{t=1}^{N_p}$ y $\{\mathbf{x}_t^{a,(i)}\}_{t=1}^{N_p}$ respectivamente, centrados en su media.

Algoritmo 4: RTS EnKS

Definir: $\{\mathbf{x}_T^{s,(i)}\}_{i=1}^{N_p} = \{\mathbf{x}_T^{a,(i)}\}_{i=1}^{N_p}$
for $t = T-1, ..., 1$ **do**
  $\mathbf{K}_t^s = \mathbf{S}_t^a((\mathbf{S}_{t+1}^f)^T\mathbf{S}_{t+1}^f)^{-1}(\mathbf{S}_{t+1}^f)^T$
  for $i = 1, ..., N_p$ **do**
    $\mathbf{x}_t^{s,(i)} = \mathbf{x}_t^{a,(i)} + \mathbf{K}_t^s(\mathbf{x}_{t+1}^{s,(i)} - \mathbf{x}_{t+1}^{f,(i)})$
  end
end

El EnKF que presentamos es solamente una implementación sencilla de esta idea pero la realidad es que existe una gran familia de filtros de Kalman por ensambles desarrollados para diferentes tipos de problemas. Los filtros de raíz cuadrada (EnSRF) evitan la necesidad de perturbar las observaciones, lo cual puede causar ruido estocástico indeseado y una subestimación de las covarianzas (Anderson, 2001; Whitaker y Hamill, 2002). Los filtros transformados (ETKF) operan sobre el espacio generado por los miembros del ensamble y consiguen una mejor eficiencia cuando el numero de miembros del ensamble es mucho menor que la dimensión del estado (Bishop et al., 2001).

En general, los EnKFs son métodos robustos, relativamente fáciles de comprender e implementar y que han sido utilizados en una amplia variedad de problemas. Debido a su popularidad, existen muchas modificaciones para mitigar los problemas más comunes que suele presentar. Los EnKFs son especialmente aptos para problemas de gran dimensionalidad ya que pueden representar distribuciones en

este tipo de espacios con una cantidad relativamente chica de miembros de ensamble. Además, se pueden optimizar numéricamente y tienen potencial para ser implementados de manera paralela. A pesar de que pueden fallar en modelos altamente no lineales que resulten en pronósticos no Gaussianos, en general pueden dar buena performance incluso en presencia de no linealidades. Esta familia de filtros ha disputado con la técnica variacional 4DVAR como método de asimilación de datos operacional en predicción meteorológica numérica.

Inflación y localización

Cuando los EnKFs son utilizados en escenarios de alta dimensionalidad surgen una variedad de problemas. Por un lado, si la cantidad de miembros de ensamble resulta pequeña en comparación con la dimensión de las variables de estado, el error de muestreo asociado puede resultar en que la matriz de covarianza predictiva $\widehat{\mathbf{P}}_t^f$ esté subestimada o que tenga deficiencias de rango (Miyoshi, 2011). Esta situación puede estar incluso profundizada por una mala especificación del error de modelo. Esto puede provocar que la dispersión del ensamble sea muy compacta (esto se suele llamar colapso del ensamble) lo cual da como resultado que la certidumbre sobre la predicción sea mayor de lo que debería ser y por lo tanto la información observacional sea menos considerada por el sistema de asimilación de datos. Esto es un problema que se retroalimenta en los sucesivos ciclos: mientras menor es la varianza del *prior*, menor la del análisis la cual, a su vez, influye en la predicción de la varianza del próximo ciclo que resulta siendo subestimada. Como resultado, el filtro comienza a ignorar las observaciones y diverge de la trayectoria que debería estimar. Un paliativo a esta situación es la llamada inflación de covarianza que consiste en agrandar artificialmente la covarianza predictiva (Anderson y Anderson, 1999). Existen diversas implementaciones de esta idea pero distinguimos la inflación aditiva que consiste en añadir ruido a los miembros del ensamble y la inflación multiplicativa que en general es implementada multiplicando la matriz de covarianzas (o las desviaciones de los miembros del ensamble respecto a su media) por un escalar mayor a 1. Para mejorar la calidad de la matriz de covarianzas predictiva, además se suele usar localización. Esta metodología consiste en reducir el valor de los términos fuera de la diagonal que correspondan a covarianzas de variables lejanas espacialmente y que por lo tanto no deberían estar muy correlacionadas. Esto se hace porque un número bajo de miembros de ensamble puede provocar correlaciones espurias que deterioren la calidad de la matriz de covarianzas (Hamill, Whitaker y Snyder, 2001). Utilizando localización se puede mitigar la deficiencia de rango y el error de muestreo.

Ejemplo Lorenz-63 Consideramos aquí el modelo Lorenz-63 descrito por las Ecuaciones 2.6. Generamos una trayectoria verdadera, a partir de la cual muestreamos observaciones sintéticas y luego, para intentar estimar el estado, utilizamos el EnKF, el filtro de partículas bootstrap y el VMPF. En la Figura 2.5 se muestran las estimaciones para la primer variable x con los tres métodos. Utilizamos solamente 2 partículas y se puede ver que el VMPF logra estimar las trayectorias reales, el EnKF se desincroniza en ciertos sectores de la ventana de tiempo y el filtro de partículas bootstrap diverge completamente luego de unos 30 ciclos de asimilación. Es inusual utilizar estas técnicas con tan pocas partículas pero el experimento pone en evidencia que el VMPF y el EnKF utilizan las partículas con mayor eficiencia que el filtro de partículas bootstrap. También se aprecia que en el EnKF y VMPF estas son mucho más próximas a la media mientras que las partículas del filtro bootstrap, al divergir

toman trayectorias que prácticamente ignoran a las observaciones. En la Figura 2.6 podemos ver la raíz del error cuadrático medio (RMSE) y la varianza de los ensambles para distinto número de partículas. Notablemente, se necesitan muchas más partículas para el filtro bootstrap para lograr una performance similar a la de las otras dos metodologías. Esto se debe en parte a que los otros dos métodos transforman el ensamble de pronóstico hacia zonas de mayor verosimilitud mientras que el filtro bootstrap se limita a seleccionar y replicar las partículas del pronóstico que resulten más verosímiles.

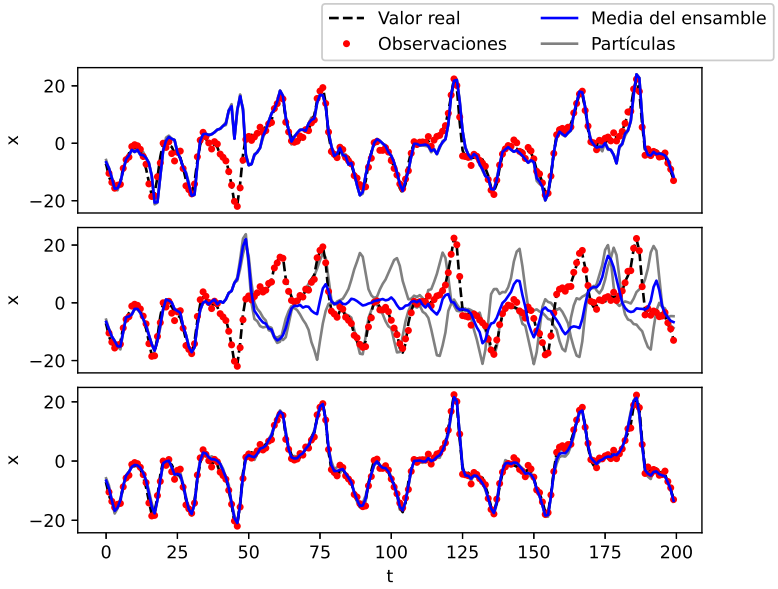

Figura 2.5: Trayectorias reales y estimadas de x mediante EnKF (arriba) filtro de partículas bootstrap (medio) y VMPF (abajo)

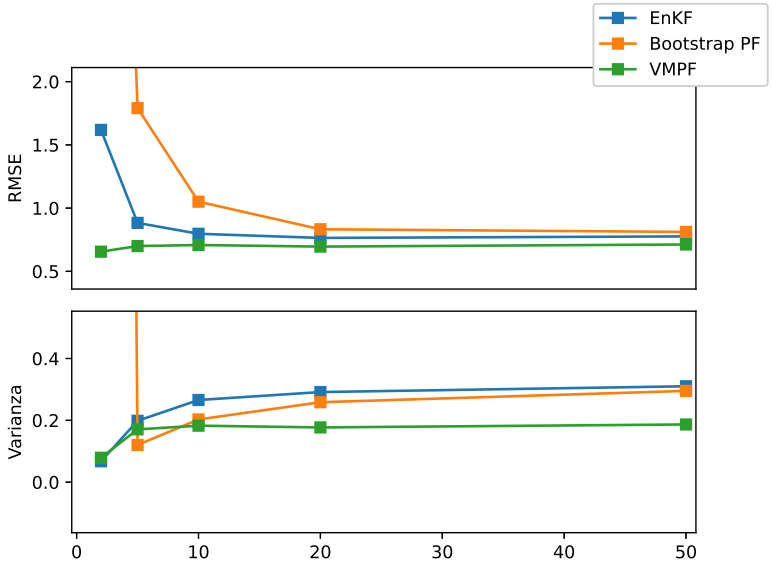

Figura 2.6: RMSE y varianza para los tres métodos

Capítulo 3

Tratamiento de errores

3.1 Error observacional y de modelo

Más allá de los desafíos y dificultades de implementación de las técnicas de asimilación de datos discutidas en el capítulo anterior, la performance de éstas depende crucialmente de la especificación del error de modelo y el observacional. Tanto la predicción de las variables de estado en el tiempo como el proceso observacional tienen fuentes de incerteza (Dee, 1995; Dee y Da Silva, 1999; Tandeo, Ailliot et al., 2020). De hecho, las metodologías de asimilación de datos hacen uso de nuestro conocimiento sobre estas incertidumbres para ponderar entre la información que brinda el pronóstico producido por el modelo y la observación. Veremos que una representación incorrecta de estos errores, y en particular de la razón entre estos, puede dar lugar a un exceso de confianza en los pronósticos o en las observaciones. Esto puede ocasionar que se degrade la calidad de las estimaciones de las distribuciones filtrantes y potencialmente que el filtro se desicronice de la trayectoria subyacente que se intenta inferir.

Las variables de estado evolucionan en el tiempo mediante la aplicación del modelo $\mathcal{M}_t$, Ecuación 2.1, el cual se diseña para representar en forma simplificada la dinámica del proceso subyacente. Por supuesto, estos modelos para sistemas complejos constituyen una representación imperfecta de la realidad que buscan describir. En lo que llamamos error de modelo, no sólo incluimos este error de representatividad sino también los provenientes de aproximaciones para simplificar el cómputo, errores numéricos y posiblemente la incerteza proveniente del desconocimiento de valores exactos de la parametrización de $\mathcal{M}_t$. Llamaremos laxamente $\mathbf{Q}$ al error de modelo usando la notación en 2.10 que corresponde a la habitual hipótesis en que lo consideremos aditivo, Gaussiano e insesgado. En forma estricta $\mathbf{Q}$ es la covarianza de la distribución de transición, la cual bajo las hipótesis mencionadas, la determina de manera única. Por otro lado, el error observacional comprende al error de representatividad del operador $\mathcal{H}_t$, las imperfecciones en su especificación y las incertezas intrínsecas de los instrumentos de medición. Análogamente al error de modelo usaremos $\mathbf{R}$ para referirnos al error observacional. También asumimos para este que es aditivo, Gaussiano e insesgado. Tenemos entonces que $\mathbf{Q}$ y $\mathbf{R}$ acumulan el error de diversas fuentes de incerteza y que además en ciertos casos, como el conocimiento incompleto sobre los fenómenos que modelamos, no disponemos de una cuantificación de estas incertidumbres.

Para ilustrar la importancia de especificar correctamente $\mathbf{Q}$ y $\mathbf{R}$ consideremos que tenemos el modelo del oscilador armónico introducido en 2.2, totalmente observado con error de modelo $\mathbf{Q} = \sigma_Q^2\mathbf{I}$ y error observacional $\mathbf{R} = \sigma_R^2\mathbf{I}$. Con esta configuración generamos una trayectoria real y sus respectivas observaciones. Luego asimilamos estas observaciones con el EnKF para recuperar el estado real, pero para ello utilizaremos distintos valores de σ_Q^2 y σ_R^2. Para evaluar la performance del EnKF

en cada repetición, consideraremos dos métricas: la raíz del error cuadrático medio (RMSE) y la cobertura. El RMSE nos informa cuan cerca esta la media de la estimación de la trayectoria real. Por otro lado la cobertura indica en que porcentaje la trayectoria real está incluida en las correspondientes bandas de confianza. Consideraremos un nivel de confianza del 95% por lo que valores mayores a este indican una sobreestimación de la dispersión de la distribución filtrante mientras que valores menores corresponden a una subestimación. En la Figura 3.1 podemos ver las estimaciones del EnKF para distintas configuraciones de $\mathbf{Q}$ y $\mathbf{R}$ (los valores reales están indicados con $\mathbf{Q}_t$ y $\mathbf{R}_t$) junto con las métricas obtenidas. En el caso en que se utiliza un $\mathbf{Q} < \mathbf{Q}_t$ y un $\mathbf{R} > \mathbf{R}_t$ se tiene que la trayectoria estimada es suave, porque al subestimar el error de modelo y sobreestimar el observacional se produce el efecto de que el sistema de asimilación considera más precisos a los pronósticos que a las observaciones y por lo tanto las trayectorias prácticamente no son corregidas y se asemejan a las del modelo. Este efecto se puede ver como un subajuste (*underfit*) a las observaciones y resulta en un RMSE alto; el comportamiento del filtro en esta situación en que las trayectorias prácticamente ignoran a las observaciones suele llamarse divergencia del filtro. En el caso contrario, en que $\mathbf{Q} > \mathbf{Q}_t$ y $\mathbf{R} < \mathbf{R}_t$, se tiene un efecto de sobreajuste (*overfit*) a las observaciones puesto que se está subestimando su error y por lo tanto las estimaciones se acercan demasiado a ellas dando lugar a trayectorias mucho menos suaves. El caso en el que se utilicen errores de modelo y observacional menores a los reales se tiene que el RMSE es pequeño pero evidentemente la dispersión está subestimada. Por otro lado si ambos errores son mayores que los verdaderos se tiene que la dispersión está sobreestimada. El caso en que se usan los errores reales corresponde entonces a una solución de compromiso entre una correcta cobertura y un RMSE no demasiado grande. Podemos ver en los mapas de calor de la Figura 3.2 que es importante la razón entre $\mathbf{Q}$ y $\mathbf{R}$: mientras este cociente es similar al cociente real entre $\mathbf{Q}_t$ y $\mathbf{R}_t$ se tendrá una estimación aproximadamente buena de la media, es decir bajo RMSE. Sin embargo la subestimación conjunta de ambos errores (manteniendo la razón entre ellos) produce una mala cobertura (subestimación de la dispersión) y lo contrario sucede con la sobreestimación conjunta. También se hace evidente que en la zona de *overfit* (esquina inferior izquierda del panel de la derecha) se produce menos RMSE que la zona de *underfit* (esquina superior derecha). Esto es debido a que cuando hay *overfit* las estimaciones tienden a interpolar las observaciones y por lo tanto se mantienen relativamente en sincronía con las trayectorias reales pero en el caso de *underfit* las estimaciones no se ven influenciadas por las observaciones y por lo tanto resultan también independientes de las trayectorias reales. Este análisis, desarrollado en mayor profundidad en Tandeo, Ailliot et al., 2020, pone en evidencia la importancia de la estimación de estas incertezas, en particular de su especificación conjunta, para maximizar la performance de las técnicas de asimilación de datos.

Como hemos visto en la Sección 2.4.2, los métodos basados en ensambles tienen tendencia a colapsar debido a errores de muestreo. El error de modelo cobra una relevancia especial pues está asociada a la dispersión de las partículas del pronóstico y por lo tanto es importante especificarlo correctamente para un buen desempeño del sistema de asimilación. Por su parte, en los filtros de partículas, el error de modelo se relaciona a la incerteza de cada partícula. Muchos filtros de partículas modernos buscan mejorar el muestreo llevando a las partículas a regiones de alta verosimilitud como por ejemplo los filtros de partículas implícitos (Atkins et al., 2013; Chorin y Tu, 2009; Zhu et al., 2016), los filtros de flujos de partículas temperados (Daum y Huang, 2009) o el filtro de partículas con mapeo variacional (Pulido y van Leeuwen, 2019). Estos suponen el conocimiento de $\mathbf{Q}$ y por lo tanto se hace relevante poder

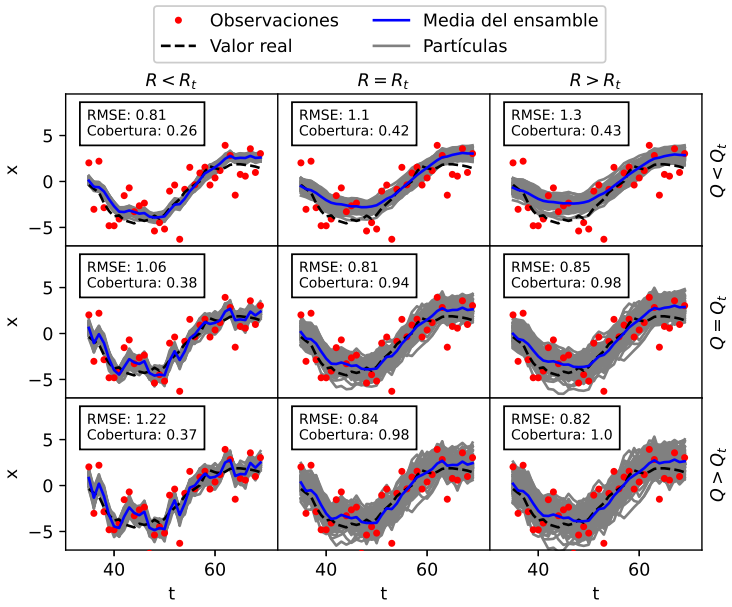

Figura 3.1: Posición x del oscilador armónico y estimaciones del EnKF para diferentes configuraciones de $\mathbf{Q}$ y $\mathbf{R}$.

acoplarlos con una metodología de estimación de dichos errores.

Se han desarrollado una gran cantidad de métodos para estimar estos errores. En Stroud et al., 2018 se apunta a maximizar la verosimilitud de las innovaciones (la diferencia entre la observación y el pronóstico mapeado al espacio observacional) utilizando inferencia Bayesiana. En otros trabajos se utilizan las covarianzas cruzadas entre innovaciones sucesivas para producir estimaciones de $\mathbf{Q}$ y $\mathbf{R}$ (Ver por ejemplo el trabajo seminal de Mehra, 1970 y una adaptación moderna basada en esta en Berry y Sauer, 2013). En el trabajo de Desroziers et al., 2005 se definen estadísticos de diagnóstico basados en las innovaciones que pueden ser utilizados para obtener coeficientes de inflación adaptativos (Li, Kalnay et al., 2009). Notemos que la inflación puede ser vista como un método de estimación del error de modelo puesto que da cuenta de la necesidad de ajustar la incertidumbre de los pronósticos. Sin embargo hay que notar que, por ser multiplicativa, amplificará la covarianza muestral sobretodo en las direcciones en que el pronóstico ya tenga mayor dispersión (Hamill y Whitaker, 2005). Otra aproximación al problema, sobre la que nos centraremos aquí, es la maximización de la verosimilitud total a través del algoritmo EM (*expectation-maximization*, Dempster et al., 1977). Este consiste en la realización iterativa de dos pasos: uno en que se computa una esperanza condicional (*E-step*) y otro en que esta esperanza se maximiza respecto al parámetro (*M-step*). El método fue acoplado con éxito al filtro de Kalman tradicional para estimar $\mathbf{Q}$ y $\mathbf{R}$ en Shumway y Stoffer, 1982 y con posteridad al filtro de Kalman por ensambles combinado con un suavizador de Kalman por ensambles (ver por ejemplo Dreano et al., 2017). Un buen compendio de todas estas técnicas se puede encontrar en Tandeo, Ailliot et al., 2020.

Dentro de toda la variedad de métodos para la estimación de errores observacionales y de modelo distinguimos los métodos *offline* de los *online*. Los primeros toman una ventana de observaciones $\mathbf{y}_{1:T}$ y dan en base a estas una única estimación para $\mathbf{Q}$ y $\mathbf{R}$ para todos los tiempos $t = 1, ..., T$. El algoritmo EM es usualmente implementado de esta manera, utilizando un lote (*batch*) de observaciones (tal es el caso en Dreano et al., 2017; Pulido, Tandeo et al., 2018; Tandeo, Pulido et al., 2015) sobre el que se aplica el EnKF en combinación con el EnKS. Este procedimiento se adaptó

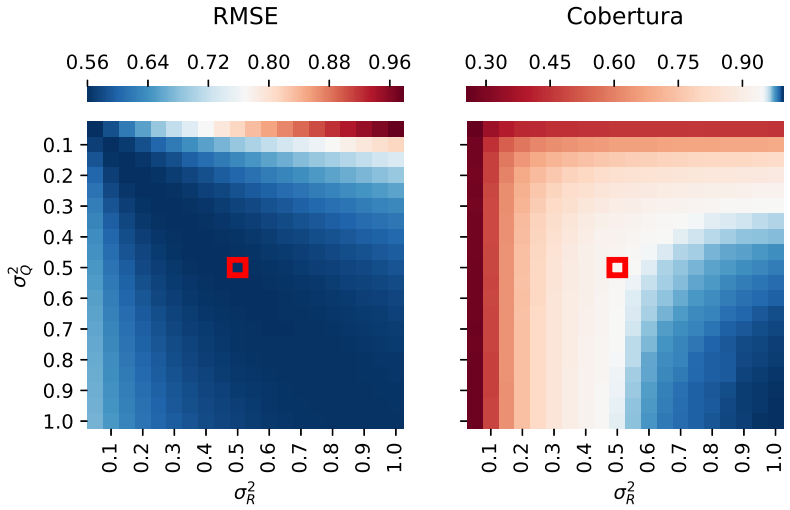

Figura 3.2: RMSE y cobertura producidas por el EnKF en la estimación del oscilador armónico para para diferentes configuraciones de **Q** y **R**. Los valores reales se indican con un recuadro azul. La barra de colores para la cobertura tiene un escalado lineal para que el centro (color blanco) esté en el valor óptimo de cobertura de 95%.

para filtros de partículas en Lucini et al., 2021 sorteando la necesidad de utilizar un suavizador de partículas. En muchas aplicaciones no se utilizan suavizadores porque sólo hay interés en las distribuciones filtrantes y pronósticos y se implementan alternando predicción con análisis. Esto ahorra el costo computacional del suavizado y el almacenamiento de todas las estimaciones anteriores necesarias para el suavizador. En estos escenarios es impráctica o inviable la aplicación de métodos de estimación *offline* y se hace necesario utilizar técnicas *online* (también llamadas secuenciales o adaptativas). Estas producen estimaciones de **Q** y **R** de manera secuencial, es decir en cada ciclo de asimilación y utilizando la información de la observación que está siendo procesada (y no de todo el lote de observaciones de manera simultánea). El trabajo de Neal y Hinton, 1998 es de gran relevancia en el área: discute varias adaptaciones para el algortimo EM tradicional, con realizaciones parciales de los *E-step* y *M-step* pero también de una versión que permite la incorporación secuencial de las observaciones a las estimaciones de los parámetros. En el contexto de modelos de Markov escondidos, podemos mencionar el algoritmo propuesto en Cappé, 2009 que implementa ideas de EM secuencial acoplados a filtros de partículas y las implementaciones de Andrieu y Doucet, 2003 que utilizan pseudo-verosimilitudes basadas en mini-lotes de datos. También es necesario mencionar que existen implementaciones *online* para estimación de errores que no están basadas en EM como por ejemplo la que se puede encontrar en Berry y Sauer, 2013.

Para esta tesis, desarrollamos un nuevo método *online* de estimación de error de modelo y observacional basado en EM compatible con filtros de partículas y con EnKFs. La técnica combina las ideas del *batch* EM en la versión de Dreano et al., 2017 con las ideas expuestas por Cappé, 2009 y Andrieu y Doucet, 2003 y el resultado está publicado en Cocucci, Pulido, Lucini et al., 2021. Para dar una derivación del método desarrollaremos el algoritmo EM tradicional por lotes en 3.3.1 y luego haremos la deducción teórica para adaptar al algoritmo EM a un esquema secuencial en 3.3.2. Sin embargo, antes de introducir estos métodos haremos la relevante mención de la técnica de estado aumentado. Esta es una metodología muy sencilla de aplicar y que se suele utilizar para la estimación de parámetros "físicos" del modelo $\mathcal{M}_t$ y que sin

embargo falla para la estimación de errores exponiendo la necesidad de tratar con métodos más involucrados.

3.2 Estado aumentado

En la sección anterior se discutió la relevancia de utilizar estimaciones apropiadas de los errores involucrados. Los parámetros que codifican a estas incertezas suelen ser llamados parámetros "estocásticos". Por otro lado, distinguimos a los parámetros específicos al modelo transicional $\mathcal{M}_t$ los cuales suelen ser llamados parámetros "determinísticos" o "físicos" ya que usualmente son cantidades interpretables como parte de la dinámica subyacente de las variables de estado $\mathbf{x}_t$. Es normal que no se cuente con un parametrización precisa del modelo de transición y por lo tanto se requiera de técnicas que permitan estimarlo a partir de mediciones del sistema. En parte, y como fue mencionado en la sección anterior, se puede dar cuenta de la imperfección en la parametrización de $\mathcal{M}_t$ a través del error de modelo y delegar a la estimación de este las incertezas de los parámetros determinísticos. Sin embargo, con la técnica conocida como "estado aumentado", es posible estimarlos individualmente y de esta manera calibrar el modelo. Esta consiste en incorporar los parámetros a las variables de estado e interpretarlas como cantidades no observadas del sistema. Si llamamos $\boldsymbol{\theta}_t$ a los parámetros que queremos estimar, construimos entonces el estado aumentado $\tilde{\mathbf{x}}_t = (\mathbf{x}_t, \boldsymbol{\theta}_t)$ (notemos la subindexación t que incluimos porque este método admite que los parámetros varíen en el tiempo). Para poder implementar esta idea es necesario extender $\mathbf{H}_t$ para que interprete a los parámetros como variables no observadas y a $\mathcal{M}_t$ para que actúe sobre estos.

Las técnicas de asimilación de datos pueden inferir sobre variables no observadas ya que la asimilación captura las correlaciones entre estas y las observaciones. En el caso de que esta correlación sea muy débil, el análisis será conservador respecto a la variable no observada que permanecerá cerca del pronóstico. Por lo tanto, este comportamiento se replica para los parámetros en estado aumentado y, en el caso que las correlaciones mencionadas sean los suficientemente fuertes, se podrán obtener estimaciones para los parámetros. Además, como estas estimaciones son secuenciales y siguen la lógica "pronóstico-análisis" como el resto de las variables de estado, es posible estimar parámetros con variación temporal dando lugar a un sistema que se auto-calibra (Annan y Hargreaves, 2004; Ruiz et al., 2013a). Sin embargo, hay que mencionar que, como los parámetros son utilizados en el modelo para el paso de tiempo subsiguiente, si los cambios en el parámetro son muy bruscos el sistema tardará en capturarlos, de manera que la adaptabilidad del método está sujeta a que las variaciones temporales de los parámetros sean lo suficientemente lentas como para que el sistema pueda asimilarlas.

Para la extensión de $\mathcal{M}_t$ sobre los parámetros es común considerar que actúa como la identidad sobre los parámetros considerando sobre estos una hipótesis de persistencia. Esto significa que si $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ entonces $\mathcal{M}_t|_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \boldsymbol{\theta}$. Sin embargo, esto puede causar que las estimaciones se limiten a los valores del *prior* por lo que es habitual incorporar una caminata aleatoria Gaussiana $\mathcal{M}_t|_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \boldsymbol{\theta} + \boldsymbol{\epsilon}_t$ con $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$. Esto contribuye a que el pronóstico de los parámetros consiga una mejor exploración del espacio paramétrico. El valor de $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ cuantifica la magnitud de los pasos de la caminata aleatoria y se constituye como un hiperparámetro que se puede calibrar para mejorar la performance del sistema de asimilación. Existen métodos para elegir el valor de $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ de manera dinámica (Doucet et al., 2001). El

método de estado aumentado también permite modelar una dinámica más compleja para la evolución de los parámetros si esto fuera necesario.

En la Figura 3.3 podemos ver un experimento usando el EnKF en el modelo Lorenz-63 utilizando estado aumentado para estimar ρ. Consideramos que el valor verdadero de este parámetro varía en el tiempo de acuerdo a una función sinusoidal. El sistema es capaz de capturar estos cambios y las estimaciones sincronizan con el valor real del parámetro luego de unas pocas iteraciones. La varianza inicial del ensamble se eligió relativamente grande para una mejor exploración del espacio paramétrico.

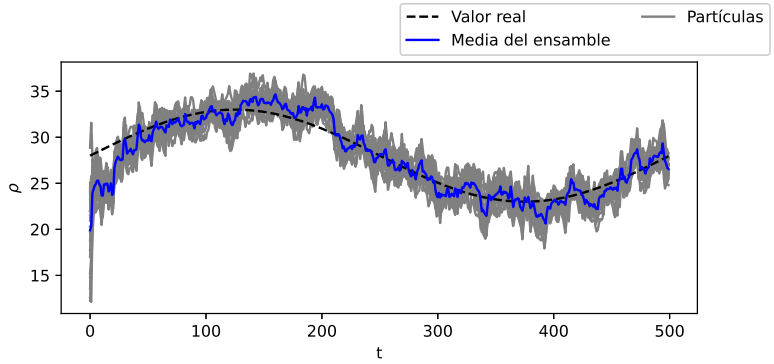

Figura 3.3: Estimación del parámetro ρ del modelo Lorenz-63 mediante estado aumentado utilizando el EnKF

La técnica de estado aumentado suele ser adecuada para muchos parámetros determinísticos. Sin embargo, no da buenos resultados para la estimación de parámetros estocásticos debido a la falta de correlación entre estos y la información observacional. En DelSole y Yang, 2010 se puede encontrar una definición algo más precisa de parámetros determinísticos y estocásticos así como una justificación más completa de por qué aumentar el estado con parámetros estocásticos no puede dar buenas estimaciones.

3.3 Algoritmo EM

El algoritmo EM se utiliza para obtener estimadores de máxima verosimilitud en sistemas parcialmente observados. Es en realidad una metodología general y no una solución *off-the-shelf*. Su aplicación más conocida es en el contexto de aprendizaje no supervisado para hacer *clustering* modelando el problema con una mezcla de Gaussianas (Bishop, 2006) pero tiene una gran diversidad de utilidades. Comenzaremos dando su forma general, luego su aplicación en lotes para estimación de matrices de covarianzas en *state-space models* con error aditivo Gaussiano y finalmente su adaptación *online*.

El algoritmo EM se aplica en el contexto de un modelo probabilístico en el que contamos con una variable observada $\mathbf{y}$, variables no observadas $\mathbf{x}$ y parámetros $\boldsymbol{\theta}$ con lo que se constituye la probabilidad conjunta $p(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})$. El objetivo es utilizar datos $\mathbf{y}$ para estimar el parámetro $\boldsymbol{\theta}$ mediante la maximización de la verosimilitud

$p(\mathbf{y};\boldsymbol{\theta})$ o equivalentemente, de su logaritmo. Si dotamos a las variables no observadas de una distribución *a priori* $q(\mathbf{x})$ arbitraria podemos obtener la expresión:

$$\log p(\mathbf{y};\boldsymbol{\theta}) = \log \int p(\mathbf{x},\mathbf{y};\boldsymbol{\theta}) d\mathbf{x} \tag{3.1}$$

$$= \underbrace{\int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{y};\boldsymbol{\theta})} d\mathbf{x}}_{KL(\int q(\mathbf{x}) \| p(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}))} + \underbrace{\int q(\mathbf{x}) \log \frac{p(\mathbf{x},\mathbf{y};\boldsymbol{\theta})}{q(\mathbf{x})} d\mathbf{x}}_{\mathcal{L}(q,\boldsymbol{\theta})} \tag{3.2}$$

donde KL es la divergencia de Kullback-Leibler y $\mathcal{L}$ es llamada ELBO (*evidence lower bound*). Es común interpretar a KL como una "distancia" entre probabilidades y de hecho, cumple que $KL(q\|p) \geq 0$ y se anula sí y sólo si $p = q$ en casi todo punto [1]. Al ser KL mayor o igual a 0, esto significa que $\log p(\mathbf{y};\boldsymbol{\theta}) \geq \mathcal{L}(q,\boldsymbol{\theta})$, es decir que la ELBO es una cota inferior de la log-verosimilitud.

El algoritmo EM provee estimaciones $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, ...$ tales que $\log p(\mathbf{y};\boldsymbol{\theta}_{t+1}) \geq \log p(\mathbf{y};\boldsymbol{\theta}_t)$ que convergen a un máximo local de la verosimilitud (Wu, 1983). Como dado un q fijo, la $\mathcal{L}(q,\boldsymbol{\theta})$ es una cota inferior de $\log p(\mathbf{y};\boldsymbol{\theta})$ para todo $\boldsymbol{\theta}$, entonces la idea es maximizar $\mathcal{L}(q,\boldsymbol{\theta})$ primero respecto a q y luego respecto a $\boldsymbol{\theta}$. Supongamos que ya contamos con la estimación de la t-ésima iteración, $\boldsymbol{\theta}_t$. Si queremos obtener $q = \underset{q}{\operatorname{argmax}}\, \mathcal{L}(q,\boldsymbol{\theta}_t)$, notemos que la igualdad 3.2 se satisface para todo q por lo que debemos elegir el valor que anule a la divergencia de Kullback-Leibler, es decir $q = p(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}_t)$. Luego, dejamos fijo q y elegimos $\theta_{t+1} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathcal{L}(q,\boldsymbol{\theta})$. De esta manera obtendremos que,

$$\begin{aligned}
\log p(\mathbf{y};\boldsymbol{\theta}_t) &= \mathcal{L}(p(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}_t),\boldsymbol{\theta}_t) + \overbrace{KL(p(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}_t)\|p(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}_t))}^{0} \\
&= \mathcal{L}(p(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}_t),\boldsymbol{\theta}_t) \\
&\leq \mathcal{L}(p(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}_t),\boldsymbol{\theta}_{t+1}) \\
&\leq \log p(\mathbf{y};\boldsymbol{\theta}_{t+1})
\end{aligned}$$

y por lo tanto que las estimaciones producidas incrementan la verosimilitud.

Notemos además que una vez elegido q tal que anule a la divergencia de Kullback-Leibler en la iteración t obtenemos que:

$$\begin{aligned}
\mathcal{L}(p(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}_t),\boldsymbol{\theta}) &= \int p(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}_t) \log \frac{p(\mathbf{x},\mathbf{y};\boldsymbol{\theta})}{p(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}_t)} d\mathbf{x} \\
&= \int p(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}_t) \log p(\mathbf{x},\mathbf{y};\boldsymbol{\theta}) d\mathbf{x} - \int p(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}_t) \log p(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}_t) d\mathbf{x} \\
&\propto_{\theta} \int p(\mathbf{x}|\mathbf{y};\boldsymbol{\theta}_t) \log p(\mathbf{x},\mathbf{y};\boldsymbol{\theta}) d\mathbf{x} \\
&\doteq E_{\boldsymbol{\theta}_t}[\log p(\mathbf{x},\mathbf{y};\boldsymbol{\theta})|\mathbf{y}]
\end{aligned}$$

es decir que la ELBO se puede expresar como una esperanza condicional una vez que elegimos q que maximiza a $\mathcal{L}$ (notar que hemos introducido la notación $E[\cdot|\cdot]$ que enfatiza el condicionamiento). Debido a esto, la maximización sobre q recibe el nombre de *E-step*. Por otra parte el *M-step* corresponde a la maximización sobre $\boldsymbol{\theta}$. El procedimiento admite entonces la caracterización que presentada en el Algoritmo

[1] Notar que la KL no es simétrica por lo que no es un distancia propiamente dicha. Aún así, existen formas naturales y sencillas de simetrizarla

5 en el que suponemos que se realizan una cantidad prefijada N_{it} de iteraciones, aunque también es posible usar otros criterios de finalización.

Algoritmo 5: EM general

Elegir valor inicial $\boldsymbol{\theta}_0$:
for $t = 0, 1, ..., N_{it}$ **do**
  E-step:
    Computar $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}_t) = E_{\boldsymbol{\theta}_t}[\log p(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})|\mathbf{y}]$
  M-step:
    $\boldsymbol{\theta}_{t+1} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$
end

La metodología que presentamos tiene la conveniencia de incrementar (o mantener) la verosimilitud en cada paso, sin embargo no garantiza que el máximo encontrado sea un máximo global de la verosimilitud. Cuando la probabilidad conjunta de las variables observadas y no observadas pertenece a la familia exponencial, tenemos simplificaciones importantes en el cómputo y el máximo puede determinarse en forma analítica; esto no siempre es el caso y existen variantes del algoritmo EM que consideran hacer una maximización parcial el *M-step* (EM generalizado). Otra generalización consiste en una optimización parcial en la elección de q en el *E-step* (EM incremental, Neal y Hinton, 1998) que se implementa mediante la incorporación secuencial de las observaciones. Este método ayuda a una convergencia más rápida del algoritmo, que de otro modo tiene una convergencia en muchos casos lenta. Además, es el punto de partida para las versiones *online* del EM que veremos más adelante.

Ejemplo EM Para ilustrar algunas de las características del algoritmo EM consideraremos una variable unidimensional no observada $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\theta}_t, \sigma_x^2)$ y una observación que responde al modelo $\mathbf{y} = (\mathbf{x} + b)^2 + \boldsymbol{\epsilon}$ donde $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_y^2)$ y b es un escalar. Es decir que tenemos una variable no observada que depende del parámetro verdadero $\boldsymbol{\theta}_t$ y una observación que corresponde a una función cuadrática de la realización de $\mathbf{x}$ más ruido aditivo Gaussiano. Esto resulta en una log-verosimilitud bimodal como se representa en la Figura 3.4. También se pueden ver las curvas de la ELBO para dos iteraciones del EM: éstas son cotas inferiores de la log-verosimilitud y en un punto ($\boldsymbol{\theta} = \boldsymbol{\theta}_i$) son iguales. La figura muestra que cada iteración necesariamente tendrá una verosimilitud mayor o igual a las anteriores. Por otro lado quedan en evidencia dos de las debilidades del método: por un lado la convergencia puede ser lenta puesto que la maximización de la ELBO puede resultar en incrementos pequeños de la verosimilitud, y por otro lado, dependiendo de la estimación inicial $\boldsymbol{\theta}_0$, el algoritmo puede converger a un máximo local. En este ejemplo la convergencia es hacia el menor de los dos máximos, pero si la estimación inicial fuera menor al valor mínimo del valle entre estos el algoritmo convergería al estimador de máxima verosimilitud. Finalmente, notemos que la diferencia entre el máximo de la ELBO para la estimación $\boldsymbol{\theta}_i$ y el valor de la verosimilitud en ese punto es $KL(p(\mathbf{x}|\mathbf{y}; \boldsymbol{\theta}_{i-1}) \| p(\mathbf{x}|\mathbf{y}; \boldsymbol{\theta}_i))$ debido a la descomposición de la verosimilitud expresada en la Ecuación 3.2.

3.3.1 *Batch* EM

Los modelos de Markov escondidos son efectivamente sistemas parcialmente observados que en principio, admiten la aplicación del algoritmo EM. Pero además

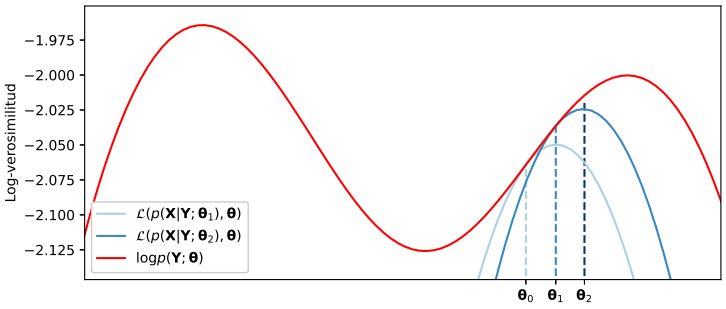

Figura 3.4: Log-verosimilitud y ELBO para dos iteraciones del algoritmo EM

la estructura Markoviana de dependencia temporal de las variables no observadas junto a la independencia condicional de las observaciones puede ser utilizada para obtener una expresión más sencilla de la ELBO. Haremos ahora la suposición de que contamos con un modelo de Markov escondido en un intervalo de tiempos $t = 0, 1, ..., T$ para los cuales tenemos variables latentes $\mathbf{x}_{0:T}$ y observaciones $\mathbf{y}_{1:T}$. Las propiedades mencionadas sobre modelos de Markov escondidos nos permiten factorizar a la probabilidad conjunta (necesaria para computar la ELBO) como:

$$p(\mathbf{x}_{0:T}, \mathbf{y}_{1:T}; \boldsymbol{\theta}) = p(\mathbf{x}_0; \boldsymbol{\theta}) \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}; \boldsymbol{\theta}) p(\mathbf{y}_t|\mathbf{x}_t; \boldsymbol{\theta}) \tag{3.3}$$

$$= p(\mathbf{x}_0; \boldsymbol{\theta}) \prod_{t=1}^{T} p(\mathbf{x}_t, \mathbf{y}_t|\mathbf{x}_{t-1}; \boldsymbol{\theta}) \tag{3.4}$$

Para obtener la ELBO correspondiente a la i-ésima iteración del método, se elige a q como la distribución de las variables latentes condicionadas a las observaciones y usando la expresión 3.4 se tiene que:

$$\mathcal{L}(p(\mathbf{x}_{0:T}|\mathbf{y}_{1:T}; \boldsymbol{\theta}_i), \boldsymbol{\theta}) \propto_{\boldsymbol{\theta}} \sum_{t=1}^{T} \int p(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}; \boldsymbol{\theta}_i) \log p(\mathbf{x}_t, \mathbf{y}_t|\mathbf{x}_{t-1}; \boldsymbol{\theta}) d\mathbf{x}_{1:T} \tag{3.5}$$

$$= \sum_{t=1}^{T} E_{\boldsymbol{\theta}_i}[\log p(\mathbf{x}_t, \mathbf{y}_t|\mathbf{x}_{t-1}; \boldsymbol{\theta})|\mathbf{y}_{1:T}] \tag{3.6}$$

En esta expresión hemos quitado el término correspondiente a $p(\mathbf{x}_0; \boldsymbol{\theta})$ bajo la suposición de que no hay parámetros desconocidos en la distribución inicial. Esta suposición no es necesaria y de hecho en Dreano et al., 2017 se opta por estimar su media y varianza como partes del vector de parámetros $\boldsymbol{\theta}$ considerados por el algoritmo EM. Ahora haremos una suposición extra que nos permitirá obtener una forma analítica del gradiente de la ELBO: supondremos que $p(\mathbf{x}_t, \mathbf{y}_t|\mathbf{x}_{t-1}; \boldsymbol{\theta})$ pertenece a la familia exponencial. Este supuesto no es extremadamente restrictivo puesto que muchas distribuciones relevantes son de la familia exponencial incluyendo, importantemente, a la Gaussiana. En ese caso tendremos entonces que:

$$p(\mathbf{x}_t, \mathbf{y}_t|\mathbf{x}_{t-1}; \boldsymbol{\theta}) = h(\mathbf{x}_t, \mathbf{y}_t) \exp(\psi(\boldsymbol{\theta}) \cdot s(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{y}_t) - A(\boldsymbol{\theta}))$$

donde $s(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{y}_t)$ es llamado el estadístico suficiente, $\psi(\boldsymbol{\theta})$ la parametrización natural y h y A son funciones (Wasserman, 2004). El gradiente de la ELBO respecto al

parámetro se puede computar como:

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}(p(\mathbf{x}_{0:T}|\mathbf{y}_{1:T};\boldsymbol{\theta}_i),\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta})\cdot\sum_{t=1}^{T} E_{\boldsymbol{\theta}_i}[s(\mathbf{x}_{t-1},\mathbf{x}_t,\mathbf{y}_t)|\mathbf{y}_{1:T}] - T\nabla_{\boldsymbol{\theta}}A(\boldsymbol{\theta}) \tag{3.7}$$

con lo cual anulando el gradiente obtenemos la siguiente ecuación

$$\nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta})\cdot S_i - \nabla_{\boldsymbol{\theta}}A(\boldsymbol{\theta}) = 0 \tag{3.8}$$

donde usamos la nomenclatura

$$S_i = \frac{1}{T}\sum_{t=1}^{T} E_{\boldsymbol{\theta}_i}[s(\mathbf{x}_{t-1},\mathbf{x}_t,\mathbf{y}_t)|\mathbf{y}_{1:T}] \tag{3.9}$$

El valor del parámetro que cumpla con 3.8 será el que maximice la ELBO y por lo tanto el valor subsiguiente del EM, $\boldsymbol{\theta}_{i+1}$. Más precisamente, el valor que anula el gradiente es un punto crítico pero en este caso está garantizado que es un máximo debido a propiedades del Hessiano en familias exponenciales (Wainwright y Jordan, 2008). Notemos que la cantidad S_i es un promedio sobre toda la ventana temporal $t = 1,..,T$ de valores esperados de los estadísticos suficientes condicionados a *todas* las observaciones y computado con la última estimación disponible del parámetro, $\boldsymbol{\theta}_i$. El condicionamiento sobre toda la ventana de observaciones implica que el valor esperado está siendo computado utilizando las distribuciones suavizantes. El método EM que se obtiene para modelos de Markov escondidos bajo la hipótesis de familia exponencial consiste entonces en un *E-step* en el que computamos S_i (3.9) y un *M-step* en el que resolvemos la ecuación que anula el gradiente de la ELBO (3.8).

El caso Gaussiano

Ahora trataremos el caso en el que el error observacional y de modelo sean aditivos y Gaussianos, es decir que tenemos:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) \sim \mathcal{N}(\mathcal{M}_t(\mathbf{x}_{t-1}),\mathbf{Q}) \tag{3.10}$$
$$p(\mathbf{y}_t|\mathbf{x}_t) \sim \mathcal{N}(\mathcal{H}_t(\mathbf{x}_t),\mathbf{R}) \tag{3.11}$$

y buscamos estimar $\boldsymbol{\theta} = (\mathbf{Q},\mathbf{R})$, las matrices de covarianzas del error de modelo y observacional respectivamente. Notemos que, no estamos considerando que estas matrices cambien en el tiempo, es decir que suponemos que son constantes en toda la ventana temporal. Además, debido a la independencia condicional de las observaciones de los modelos de Markov escondidos, tenemos que $p(\mathbf{x}_t,\mathbf{y}_t|\mathbf{x}_{t-1}) = p(\mathbf{x}_t|\mathbf{x}_{t-1})(\mathbf{y}_t|\mathbf{x}_t)$ y como supusimos que $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ y $(\mathbf{y}_t|\mathbf{x}_t)$ son Gaussianas, esto implica que $p(\mathbf{x}_t,\mathbf{y}_t|\mathbf{x}_{t-1})$ también lo es. Por lo tanto seguimos bajo la suposición de familia exponencial que enunciamos anteriormente. La cantidad S_i, en este caso puede ser pensada como una tupla $(S_i^{\mathbf{Q}},S_i^{\mathbf{R}})$ y la podemos computar mediante las siguientes expresiones que corresponden al *E-step*:

$$S_i^{\mathbf{Q}} = \frac{1}{T}\sum_{t=1}^{T} E_{\boldsymbol{\theta}_i}[(\mathbf{x}_t - \mathcal{M}_t(\mathbf{x}_{t-1}))(\mathbf{x}_t - \mathcal{M}_t(\mathbf{x}_{t-1}))^T|\mathbf{y}_{1:T}] \tag{3.12}$$

$$S_i^{\mathbf{R}} = \frac{1}{T}\sum_{t=1}^{T} E_{\boldsymbol{\theta}_i}[(\mathbf{y}_t - \mathcal{H}_t(\mathbf{x}_t))(\mathbf{y}_t - \mathcal{H}_t(\mathbf{x}_t))^T|\mathbf{y}_{1:T}] \tag{3.13}$$

Por otro lado, la Ecuación 3.8, para el caso Gaussiano tiene como solución exactamente a la cantidad S_i, con lo cual el *M-step* no requiere ningún cómputo adicional. La verificación de que $\boldsymbol{\theta} = S_i$ anula al gradiente de la ELBO se puede encontrar en el Apéndice B.2 y la representación de una densidad Gaussiana multivariada como miembro de la familia exponencial en el Apéndice B.1.

Las Ecuaciones 3.12 y 3.13, nos dan fórmulas para computar sucesivas estimaciones de $\mathbf{Q}$ y $\mathbf{R}$ y constituyen las fórmulas principales utilizadas en Dreano et al., 2017; Pulido, Tandeo et al., 2018; Tandeo, Pulido et al., 2015. Sin embargo, requieren el cómputo de valores esperados condicionados a la totalidad de la ventana de observaciones, $\mathbf{y}_{1:T}$. Si se cuenta con una representación de partículas de las distribuciones suavizantes $p(\mathbf{x}_t|\mathbf{x}_{1:T})$ para todo t, entonces los valores esperados se pueden aproximar con estimadores de Monte Carlo. Notablemente, el EnKS es una técnica que provee estas distribuciones y que también es apropiada para sistemas con errores Gaussianos aditivos; por lo tanto es compatible con esta aplicación del EM. En el algoritmo 6 podemos encontrar la implementación de este método.

Algoritmo 6: EM-EnKS

Muestrear ensamble inicial: $\{\mathbf{x}_0^{a,(i)}\}_{i=1}^{N_p} \sim p(\mathbf{x}_0)$
Elegir valor inicial $\boldsymbol{\theta}_0 = (\mathbf{Q}_0, \mathbf{R}_0)$:
for $i = 1, ..., N_{it}$ **do**
E-step: Computar los ensambles de pronóstico, filtrantes usando EnKF con la parametrización $\boldsymbol{\theta}_{i-1}$

$$\{\mathbf{x}_t^{f,(j)}\}_{j=1}^{N_p} \sim p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) \qquad \forall t = 1, ..., T$$
$$\{\mathbf{x}_t^{a,(j)}\}_{j=1}^{N_p} \sim p(\mathbf{x}_t|\mathbf{y}_{1:t}) \qquad \forall t = 1, ..., T$$

Utilizar los ensambles de pronóstico y filtrantes para computar los suavizantes mediante EnKS:

$$\{\mathbf{x}_t^{s,(j)}\}_{j=1}^{N_p} \sim p(\mathbf{x}_t|\mathbf{y}_{1:T}) \qquad \forall t = 1, ..., T$$

$$S_i^{\mathbf{Q}} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_p}\sum_{j=1}^{N_p}(\mathbf{x}_t^{s,(j)} - \mathcal{M}_t(\mathbf{x}_{t-1}^{s,(j)}))(\mathbf{x}_t^{s,(j)} - \mathcal{M}_t(\mathbf{x}_{t-1}^{s,(j)}))^T$$

$$S_i^{\mathbf{R}} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_p}\sum_{j=1}^{N_p}(\mathbf{y}_t - \mathcal{H}_t(\mathbf{x}_t^{s,(j)}))(\mathbf{y}_t - \mathcal{H}_t(\mathbf{x}_t^{s,(j)}))^T$$

M-step:
Asignar nuevos parámetros $\boldsymbol{\theta}_i = (\mathbf{Q}_i, \mathbf{R}_i)$

$$\mathbf{Q}_i = S_i^{\mathbf{Q}}$$
$$\mathbf{R}_i = S_i^{\mathbf{R}}$$

end

Podemos ver que el algoritmo involucra, para cada iteración, procesar las iteraciones hacia adelante mediante predicción y filtrado con el EnKF, reprocesarlas hacia atrás con el EnKS y luego computar con Monte Carlo las actualizaciones de los parámetros. Las pasadas hacia adelante y hacia atrás provienen de que el EnKF y

EnKS son implementaciones del algoritmo *forward-backward*. Esto significa que para utilizar este algoritmo debemos procesar todas las observaciones reiteradas veces. Además del costo computacional, esto implica que las observaciones tienen que ser almacenadas y no se contempla una posible incorporación de nuevas observaciones, situación que sería esperable en un sistema en tiempo real. Aunque el *batch* EM combinado con EnKS es un método robusto para estimar la estructura general de $\mathbf{Q}$ y $\mathbf{R}$ su naturaleza *offline* lo puede hacer impráctico en algunas situaciones y demasiado costoso computacionalmente. Además, no siempre es posible o factible obtener valores esperados respecto a distribuciones suavizantes. Por estos motivos se han desarrollado técnicas *online* o secuenciales de estimación de parámetros estocásticos.

Ejemplo EM-EnKS

Hacemos aquí una aplicación del EM-EnKS sobre el modelo del oscilador armónico para la estimación conjunta de $\mathbf{Q}$ y $\mathbf{R}$. En la Figura 3.5 podemos ver el error cuadrático medio entre las estimaciones y el valor real de los parámetros utilizados para generar las observaciones. Además mostramos la media de la diagonal de las matrices estimadas en cada iteración. En realidad, cada entrada de las matrices está siendo estimada individualmente (salvo por los elementos simétricos respecto a la diagonal). Notamos que la convergencia es rápida en las primeras iteraciones y luego se desacelera. En la Figura 3.6 se muestra el error cuadrático medio de las variables de estado suavizadas respecto a los respectivos valores reales y la log-verosimilitud de las observaciones computada mediante la siguiente aproximación desarrollada en mayor detalle en el Apéndice B.4:

$$\log p(\mathbf{y}_{1:T}) \approx \sum_{t=1}^{T} \log \frac{1}{N_p} \sum_{j=1}^{N_p} p(\mathbf{y}_t | \mathbf{x}_t^{f,(j)}) \tag{3.14}$$

Se puede ver que mientras el RMSE disminuye, la verosimilitud aumenta. Notamos también que ambas métricas, son ruidosas debido a que la técnica de asimilación sólo aproxima mediante muestras a las distribuciones.

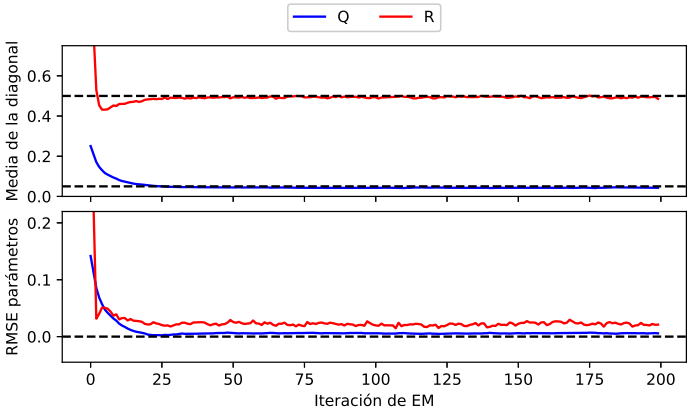

Figura 3.5: Panel superior: estimaciones de las medias de las diagonales de $\mathbf{Q}$ y $\mathbf{R}$. Las líneas intermitentes indican los valores reales. Panel inferior: errores cuadráticos medios de las estimaciones respecto a sus valores reales.

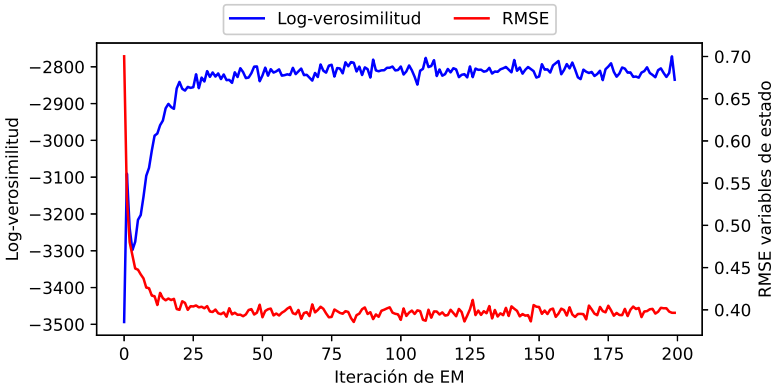

Figura 3.6: Log-verosimilitud y RMSE de las variables de estado suavizadas en cada iteración.

3.3.2 EM *online*

Aquí expondremos el algoritmo *online* basado en EM cuyo desarrollo fue publicado en Cocucci, Pulido, Lucini et al., 2021. El objetivo es obtener una técnica que actualice la estimación del parámetro con cada nueva observación de manera que se puedan descartar las observaciones anteriores que ya han sido procesadas. Si tomamos como punto de partida las Ecuaciones 3.12 y 3.13 podemos ver que, cada sumando corresponde a una observación pero que si quisiéramos agregar una observación nueva (correspondiente al tiempo $T+1$) todos estos sumandos deberían ser recomputados. Esto es debido a que los valores esperados están condicionados a toda la ventana observacional anterior, $\mathbf{y}_{1:T}$. Las nuevas distribuciones predictivas y filtrantes, $p(\mathbf{x}_{T+1}|\mathbf{y}_{1:T})$ y $p(\mathbf{x}_{T+1}|\mathbf{y}_{1:T+1})$ pueden ser obtenidas o aproximadas utilizando las anteriores distribuciones predictivas y filtrantes que no necesitan ser cambiadas por la incorporación de la nueva observación. Sin embargo, las distribuciones suavizantes $p(\mathbf{x}_t|\mathbf{y}_{1:T})$ deben ser cambiadas en cada t por las que tienen en cuenta a la nueva observación, $p(\mathbf{x}_t|\mathbf{x}_{1:T+1})$.

En este punto, dado que buscamos procesar observaciones que se hacen disponibles una por una, cambiaremos la notación de las iteraciones del EM y la haremos coincidir con la de las observaciones, puesto que queremos obtener una actualización de los parámetros por cada observación. Entonces consideramos que tenemos observaciones $\mathbf{y}_{1:T+1}$ y estimaciones de los parámetros $\boldsymbol{\theta}_0, ..., \boldsymbol{\theta}_T$, y buscaremos, a partir de esto, obtener la estimación $\boldsymbol{\theta}_{T+1}$. Más precisamente, buscaremos actualizaciones de la ELBO, es decir que dada una secuencia $S_1, ..., S_T$ buscaremos actualizar S_T para que incorpore la observación $\mathbf{y}_{T+1}$ de manera de obtener S_{T+1}. Esto es porque al hacer una extensión *online* de el *E-step* dotamos de esta propiedad a todo el algoritmo, pues el *M-step* seguirá consistiendo en solucionar 3.8 para $\boldsymbol{\theta}$ una vez computado S_{T+1}. Comenzamos entonces escribiendo la definición de S_{T+1} como en 3.9

con la nueva notación y desglosando la suma de la siguiente manera:

$$S_{T+1} = \frac{1}{T+1}\sum_{t=1}^{T+1} E_{\boldsymbol{\theta}_T}[s(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{y}_t)|\mathbf{y}_{1:T+1}] \tag{3.15}$$

$$= \frac{1}{T+1}\sum_{t=1}^{T+1} \int p(\mathbf{x}_{t-1}, \mathbf{x}_t|\mathbf{y}_{1:T+1}; \boldsymbol{\theta}_T) s(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{y}_t) d\mathbf{x}_{t-1:t} \tag{3.16}$$

$$= \frac{1}{T+1}\left(\sum_{t=1}^{T} \int p(\mathbf{x}_{t-1}, \mathbf{x}_t|\mathbf{y}_{1:T+1}; \boldsymbol{\theta}_T) s(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{y}_t) d\mathbf{x}_{t-1:t} \tag{3.17}$$

$$+ \int p(\mathbf{x}_T, \mathbf{x}_{T+1}|\mathbf{y}_{1:T+1}; \boldsymbol{\theta}_T) s(\mathbf{x}_T, \mathbf{x}_{T+1}, \mathbf{y}_{T+1}) d\mathbf{x}_{T:T+1}\right) \tag{3.18}$$

Podemos identificar entonces que los primeros T términos de la suma son similares a la cantidad S_T con la salvedad de que en el condicionamiento del valor esperado se incluye la información de la última observación. Haciendo entonces la suposición de que esta última observación no afecta significativamente a los estados anteriores y sólo influye en el último término se motiva la siguiente aproximación:

$$\widehat{S_{T+1}} = \left(1 - \frac{1}{T+1}\right)\widehat{S_T} + \frac{1}{T+1}\int p(\mathbf{x}_T, \mathbf{x}_{T+1}|\mathbf{y}_{1:T+1}; \boldsymbol{\theta}_T) s(\mathbf{x}_T, \mathbf{x}_{T+1}, \mathbf{y}_{T+1}) d\mathbf{x}_{T:T+1} \tag{3.19}$$

$$= (1 - \gamma_{T+1})\widehat{S_T} + \gamma_{T+1} E_{\boldsymbol{\theta}_T}[s(\mathbf{x}_T, \mathbf{x}_{T+1}, \mathbf{y}_{T+1})|\mathbf{y}_{1:T+1}] \tag{3.20}$$

Tenemos entonces una fórmula que nos permite computar las aproximaciones $\widehat{S}_t$ para todo t de manera recursiva en base a $\widehat{S_{t-1}}$. Para iniciar la recursión es necesario que contemos con una aproximación inicial S_0. Además introducimos γ_t que, si bien debe valer $1/t$ para satisfacer 3.20, puede ser interpretada como una tasa de aprendizaje $\gamma_t \in (0, 1)$, tomando como inspiración técnicas de aproximación estocástica LeGland y Mevel, 1997. Este parámetro va a controlar la "memoria" de los estimadores, es decir, pondera la importancia de las estimaciones anteriores respecto al nuevo término que incluye a la última observación. Como veremos luego este parámetro se puede calibrar para obtener distintos comportamientos del método en cuanto a convergencia. Notemos que con este esquema se puede flexibilizar la hipótesis del EM *batch* de que los parámetros no varían y podemos considerar casos en que los parámetros varíen lentamente en el tiempo. El método resultante tiene algunas similitudes con el propuesto en Cappé, 2009 en el que se utiliza una función auxiliar, relacionada a una forma recursiva de suavizado, que permite mantener actualizaciones de S_t. Podemos ver que, a pesar de evitar un suavizado hacia atrás hasta la primera observación, el cómputo del valor esperado en 3.20 implica un suavizado de un paso hacia atrás porque el estadístico s depende de $\mathbf{x}_T$ y el condicionamiento incluye a $\mathbf{y}_{T+1}$.

En el algoritmo 7 se esquematiza el procedimiento de manera general. Notemos que no consideramos una ventana finita de observaciones porque potencialmente se puede seguir iterando a medida que nuevos datos se hacen disponibles. Por otro lado, no hacemos suposiciones sobre los valores iniciales pero sería natural tomar S_0 y $\boldsymbol{\theta}_0$ tales que satisfagan 3.8.

De acuerdo a como se compute o aproxime el valor esperado

$$E_{\boldsymbol{\theta}_{t-1}}[s(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{y}_t)|\mathbf{y}_{1:t}] = \int p(\mathbf{x}_{t-1}, \mathbf{x}_t|\mathbf{y}_{1:t}; \boldsymbol{\theta}_{t-1}) s(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{y}_t) d\mathbf{x}_{t-1:t} \tag{3.21}$$

Algoritmo 7: EM *online*

Elegir valor inicial para el parámetro, $\boldsymbol{\theta}_0$ y el estadístico, $\widehat{S}_0$:
for $t = 1, 2, \ldots$ **do**
  E-step:
    $\widehat{S}_t = (1 - \gamma_t)\widehat{S_{t-1}} + \gamma_t E_{\boldsymbol{\theta}_{t-1}}[s(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{y}_t)|\mathbf{y}_{1:t}]$
  M-step:
  Definir $\boldsymbol{\theta}_t$ como el valor de $\boldsymbol{\theta}$ que solucione:
    $\nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}) \cdot \widehat{S}_t - \nabla_{\boldsymbol{\theta}}A(\boldsymbol{\theta}) = 0$
end

tendremos distintas implementaciones del método. En particular daremos dos posibles formas de aproximar esta integral con Monte Carlo. El primero de los métodos está basado en muestreo de importancia y está pensado para ser acoplado a filtros de partículas. La elección de la distribución de importancia evita hacer un paso de suavizado explícito. El segundo se basa en EnKF y agrega un paso hacia atrás de suavizado de manera explícita usando EnKS.

EM *online* con muestreo de importancia

Para elegir una distribución de importancia conveniente para aproximar 3.21 primero desarrollaremos $p(\mathbf{x}_{t-1}, \mathbf{x}_t|\mathbf{y}_{1:t}; \boldsymbol{\theta}_{t-1})$ de la siguiente manera, quitando la dependencia de $\boldsymbol{\theta}_{t-1}$ para mayor claridad:

$$p(\mathbf{x}_{t-1}, \mathbf{x}_t|\mathbf{y}_{1:t}) = p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})\frac{p(\mathbf{y}_t|\mathbf{x}_t)}{p(\mathbf{y}_t|\mathbf{y}_{1:t-1})} \tag{3.22}$$

En la factorización (desarrollada con mayor detalle en el Apéndice B.3) podemos reconocer al modelo de transición y observacional, a la probabilidad filtrante a tiempo $t-1$ y a la cantidad $p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$ que suele ser llamada verosimilitud marginalizada y que no depende de las variables de integración $\mathbf{x}_{t-1}$ y $\mathbf{x}_t$. Para aproximar entonces 3.21 con muestreo de importancia debemos tener muestras de ambas variables de integración. En lugar de muestrear directamente de $p(\mathbf{x}_{t-1}, \mathbf{x}_t|\mathbf{y}_{1:t})$ 3.22 sugiere que podemos muestrear de $p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$ y obtener pesos proporcionales a $p(\mathbf{y}_t|\mathbf{x}_t)$ y, mientras que nos aseguremos de normalizar los pesos, no debemos preocuparnos por la verosimilitud marginal. Esto es conveniente porque disponemos del modelo observacional, $p(\mathbf{y}_t|\mathbf{x}_t)$ para evaluar los pesos. Además, como cualquier técnica de filtrado por ensambles produce muestras de la distribución filtrante, podemos utilizar esas partículas como representación de $p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$. Asimismo, obtener muestras de $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ es equivalente a inegrar el modelo de transición sobre $\mathbf{x}_{t-1}$.

Si estamos utilizando un método por ensambles de N_p partículas, podemos utilizar nuestra muestra de la distribución filtrante,

$$\{\mathbf{x}_{t-1}^{a,(j)}\}_{j=1}^{N_p} \sim p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}; \boldsymbol{\theta}_{t-1})$$

y en base a cada partícula de esta muestra obtener otra, cuyo tamaño denominamos M_p, correspondiente a $\mathbf{x}_t$:

$$\{\mathbf{x}_t^{f,(j,l)}\}_{l=1}^{M_p} \sim p(\mathbf{x}_t|\mathbf{x}_{t-1}^{a,(j)}; \boldsymbol{\theta}_{t-1})$$

Estas últimas $N_p M_p$ partículas llevan el superíndice f pues se obtienen de la misma manera en que se obtendría un pronóstico (*forecast*) pero haciendo la salvedad de que tenemos M_p partículas por cada punto del tiempo anterior. Con estas muestras podemos ya calcular los pesos no normalizados,

$$\overline{w_{j,l}} = p(\mathbf{y}_t|\mathbf{x}_t^{(j,l)})$$

y una vez que obtenemos las versiones normalizadas, $w_{j,l}$ podemos hacer la aproximación de Monte Carlo de la integral:

$$E_{\boldsymbol{\theta}_{t-1}}[s(\mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{y}_t)|\mathbf{y}_{1:t}] \approx \sum_{j=1}^{N_p} \sum_{l=1}^{M_p} w_{j,l} s(\mathbf{x}_{t-1}^{a,(j)}, \mathbf{x}_t^{f,(j,l)}, \mathbf{y}_t)$$

Este método puede ser implementado con cualquier técnica de asimilación de datos por ensambles ya que solo necesitamos: una representación de partículas filtrante, evaluar el modelo observacional, $p(\mathbf{y}_t|\mathbf{x}_t)$, y muestrear del modelo de transición, $p(\mathbf{x}_t|\mathbf{x}_{t-1})$, es decir evolucionar el modelo hacia adelante. Por lo tanto, la metodología es compatible con EnKF y filtros de partículas. Notemos también que es posible tomar $M_p = 1$ pues esto es equivalente a muestrear $(\mathbf{x}_{t-1}, \mathbf{x}_t)$ de manera conjunta de la distribución $p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) = p(\mathbf{x}_t, \mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})$. El Algoritmo 8 especifica el método.

Algoritmo 8: EM *online* con muestreo de importancia

Muestrear partículas iniciales: $\{\mathbf{x}_0^{(j)}\}_{j=1}^{N_p} \sim p(\mathbf{x}_0)$
Elegir valor inicial para el parámetro, $\boldsymbol{\theta}_0$ y el estadístico, $\widehat{S_0}$:
for $t = 1, 2, ...$ **do**
  Calcular pesos:
  for $j = 1, ..., N_p$ **do**
    for $l = 1, ..., M_p$ **do**
      $\mathbf{x}_t^{f,(j,l)} \sim p(\mathbf{x}_t|\mathbf{x}_{t-1}^{a,(j)}; \boldsymbol{\theta}_{t-1})$
      $w_{j,l} \propto p(\mathbf{y}_t|\mathbf{x}_t^{f,(j,l)}; \boldsymbol{\theta}_{t-1})$
    end
  end
  Muestrear partículas filtrantes:
    $\{\mathbf{x}_t^{a,(j)}\}_{j=1}^{N_p} \sim p(\mathbf{x}_t|\mathbf{y}_{1:t}; \boldsymbol{\theta}_{t-1})$
  Computar $\widehat{S_t}$:
    $\widehat{S_t} = (1-\gamma_t)\widehat{S_{t-1}} + \gamma_t \sum_{j=1}^{N_p} \sum_{l=1}^{M_p} w_{j,l} s(\mathbf{x}_{t-1}^{a,(j)}, \mathbf{x}_t^{f,(j,l)}, \mathbf{y}_t)$
  Definir $\boldsymbol{\theta}_t$ como el valor de $\boldsymbol{\theta}$ que solucione:
    $\nabla_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta}) \cdot \widehat{S_t} - \nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta}) = 0$
end

EM *online* con un paso de suavizado

Una segunda propuesta para obtener una aproximación de Monte Carlo de 3.21 es muestrear directamente de $p(\mathbf{x}_{t-1}, \mathbf{x}_t|\mathbf{y}_{1:t})$ para lo cual es necesario obtener la versión suavizada del estado a tiempo $t-1$. El EnKF puede proveernos las partículas correspondientes a $p(\mathbf{x}_t|\mathbf{y}_{t-1})$ mientras que podemos obtener una muestra de $p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t})$

mediante el EnKS, presentado en el algoritmo 4. Notemos además que no es necesario iterar con el EnKS hasta la primera observación sino que basta con hacer un solo paso hacia atrás pues no estamos interesados en las partículas de tiempos anteriores a $t-1$. De esta manera, muestreamos:

$$\{\mathbf{x}_t^{a,(j)}\}_{j=1}^{N_p} \sim p(\mathbf{x}_t|\mathbf{x}_{1:t})$$
$$\{\mathbf{x}_t^{s,(j)}\}_{j=1}^{N_p} \sim p(\mathbf{x}_{t-1}|\mathbf{x}_{1:t})$$

donde las partículas suavizantes pueden ser obtenidas del ensamble de pronóstico y el filtrante utilizando las siguientes expresiones:

$$\mathbf{K}_{t-1}^s = \mathbf{S}_{t-1}^a({\mathbf{S}_t^f}^T\mathbf{S}_t^f)^{-1}{\mathbf{S}_t^f}^T$$
$$\mathbf{x}_{t-1}^{j,(s)} = \mathbf{x}_{t-1}^{j,(s)} + \mathbf{K}_{t-1}^s(\mathbf{x}_t^{j,(s)} - \mathbf{x}_t^{j,(f)})$$

donde $\mathbf{S}_t^f$ y $\mathbf{S}_t^a$ son tal como se las especifica en la Sección 2.4.2. Notemos que, cuando $\mathbf{y}_t$ es la última observación disponible, la distribución filtrante coincide con la suavizante a ese tiempo y por lo tanto $\mathbf{x}_t^{s,(j)} = \mathbf{x}_t^{a,(j)}$. La aproximación de Monte Carlo resulta entonces en:

$$E_{\boldsymbol{\theta}_{t-1}}[s(\mathbf{x}_{t-1},\mathbf{x}_t,\mathbf{y}_t)|\mathbf{y}_{1:t}] \approx \frac{1}{N_p}\sum_{j=1}^{N_p} s(\mathbf{x}_{t-1}^{s,(j)},\mathbf{x}_t^{a,(j)},\mathbf{y}_t)$$

y la metodología puede ser entonces expresada en forma algorítmica como se expone en el Algoritmo 9.

Algoritmo 9: EM *online* con suavizado de un paso

Muestrear partículas iniciales: $\{\mathbf{x}_0^{(j)}\}_{j=1}^{N_p} \sim p(\mathbf{x}_0)$
Elegir valor inicial para el parámetro, $\boldsymbol{\theta}_0$ y el estadístico, $\widehat{S_0}$:
for $t = 1, 2, \ldots$ **do**
  Computar partículas filtrantes y suavizantes utilizando EnKF+EnKS:
    $\{\mathbf{x}_t^{a,(j)}\}_{j=1}^{N_p} \sim p(\mathbf{x}_t|\mathbf{y}_{1:t};\boldsymbol{\theta}_{t-1})$
    $\{\mathbf{x}_{t-1}^{s,(j)}\}_{j=1}^{N_p} \sim p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t};\boldsymbol{\theta}_{t-1})$
  Computar $\widehat{S}_t$:
    $\widehat{S}_t = (1-\gamma_t)\widehat{S_{t-1}} + \gamma_t\frac{1}{N_p}\sum_{j=1}^{N_p} s(\mathbf{x}_{t-1}^{s,(j)},\mathbf{x}_t^{a,(j)},\mathbf{y}_t)$
  Definir $\boldsymbol{\theta}_t$ como el valor de $\boldsymbol{\theta}$ que solucione:
    $\nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta})\cdot\widehat{S}_t - \nabla_{\boldsymbol{\theta}}A(\boldsymbol{\theta}) = 0$
end

No especificamos qué implementación de EnKF o EnKS utilizamos porque potencialmente podemos elegir la que resulte más conveniente. Al estar basado en el filtro de Kalman por ensambles, se tienen los requerimientos usuales para dicho método, es decir errores aditivos Gaussianos en el modelo de transición y observacional. Si consideramos que en el caso Gaussiano, el estadístico suficiente se puede expresar como $s(\mathbf{x}_{t-1},\mathbf{x}_t,\mathbf{y}_{1:t}) = (s^Q(\mathbf{x}_{t-1},\mathbf{x}_t), s^R(\mathbf{y}_t,\mathbf{x}_t))$ donde

$$s^Q(\mathbf{x}_{t-1},\mathbf{x}_t) = (\mathbf{x}_t - \mathcal{M}_t(\mathbf{x}_{t-1}))(\mathbf{x}_t - \mathcal{M}_t(\mathbf{x}_{t-1}))^T$$
$$s^R(\mathbf{y}_t,\mathbf{x}_t) = (\mathbf{y}_t - \mathcal{Y}_t(\mathbf{x}_t))(\mathbf{y}_t - \mathcal{H}_t(\mathbf{x}_t))^T$$

podemos ver la similitud entre este método y el EM-EnKS presentado en el Algoritmo 6. Sin embargo, si aplicáramos la versión *online* a una ventana de tiempo $t = 1, ..., T$, podemos ver que este realiza aproximadamente la misma cantidad de operaciones en total que una única iteración de la versión *offline*. De hecho, el EM *online* requerirá T pasos temporales de EnKF y T pasos simples (de una sola observación) hacia atrás de EnKS: esto es equivalente al pase completo de EnKF+EnKS en la versión *batch*. Por otro lado, la cantidad de evaluaciones de s en el pase completo de la versión *online* es $N_p T$, lo cual coincide con una única iteración del *offline*. El costo computacional es entonces menor para la versión *online* y de aplicación más directa a sistemas secuenciales.

En la implementación de muestreo de importancia, los ensambles representan muestras de las distribuciones. En esta versión basada en EnKF también pero con la importante salvedad de que las muestras están computadas bajo suposiciones de Gaussianidad. Esto significa que se considera que las distribuciones pueden ser representadas solamente con los dos primeros momentos mientras que los de orden superior son ignorados.

3.3.3 EM *online*: evaluación experimental

Expondremos aquí una evaluación experimental del método en distintos escenarios de interés para aplicaciones en asimilación de datos. La configuración de los experimentos consiste en generar trayectorias reales de las variables de estado, mediante la llamada simulación real o natural, y de estas obtener observaciones sintéticas con parametrizaciones conocidas para $\mathbf{Q}$ y $\mathbf{R}$. Luego aplicaremos el EM *online* para estimar las variables de estado e importantemente la parametrización original. Sólo utilizamos como fuente de información las observaciones sintéticas, es decir que el sistema desconoce las covarianzas y la simulación natural. Utilizaremos el Algoritmo 9 con el EnKF estocástico (Algoritmo 3) y el paso de suavizado hacia atrás mediante el RTS-EnKS (Algoritmo 4) pero realizando un solo paso hacia atrás. Daremos a esta implementación el nombre OSS-EnKF (por *one step smoother*). Por otro lado, para el Algoritmo 8 utilizaremos dos implementaciones distintas, una con el EnKF estocástico la cual llamaremos IS-EnKF (por *importance sampling*) y otra con el VMPF la cual denominaremos IS-VMPF. Para esta última implementación notamos que la estimación de la matriz de covarianzas del error de modelo $\mathbf{Q}$ tiene una relevancia adicional pues utilizamos un kernel $K(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \mathbf{x}')^T \beta \mathbf{Q} (\mathbf{x} - \mathbf{x}')$, por lo que el ancho de banda del kernel dependerá de las estimaciones de $\mathbf{Q}$. Para todas las versiones del EM *online* utilizaremos una tasa de aprendizaje $\gamma_t = t^{-\alpha}$ (LeGland y Mevel, 1997) donde el valor por defecto es $\alpha = 0.6$ con la excepción de un experimento en el que evaluaremos la performance para distintos valores de α. En base a los resultados de dicho experimento es que tomamos el valor por defecto $\alpha = 0.6$.

Consistencia respecto a valores iniciales

En este experimento utilizamos el modelo Lorenz-63 y estimamos $\mathbf{Q}$ con las 3 implementaciones, considerando a $\mathbf{R}$ conocido. Esto se repite para distintas semillas de los generadores aleatorios y distintos valores iniciales del parámetro. El objetivo es evaluar la consistencia de las estimaciones y la comparabilidad de los métodos. La matriz $\mathbf{Q}$ utilizada en en la simulación natural es un múltiplo de la identidad. En la Figura 3.7 se muestran las estimaciones de la media de la diagonal de $\mathbf{Q}$, es decir la varianza media del error de modelo. Notemos que las trayectorias grises corresponden a repeticiones del experimento y no deben ser confundidas con miembros

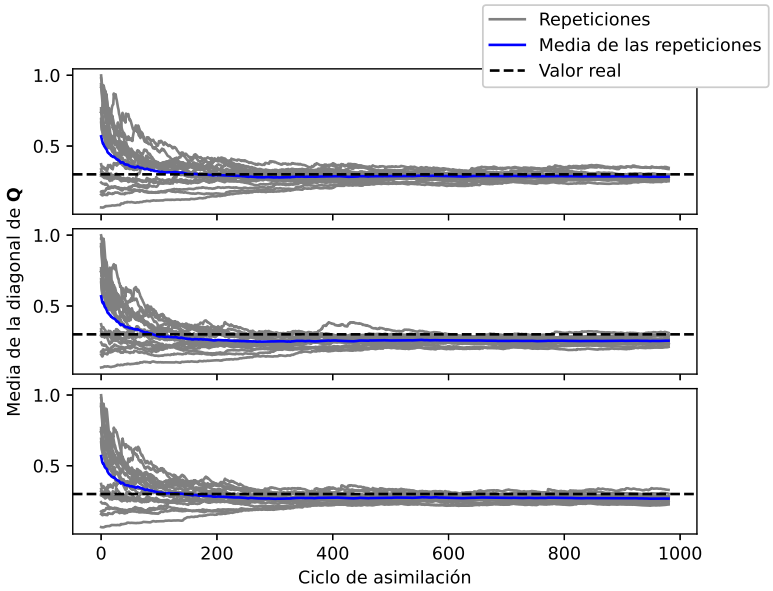

Figura 3.7: Estimaciones de la media de la diagonal de **Q** con OSS-EnKF (panel superior), IS-EnKF (panel medio) y IS-VMPF (panel inferior).

de ensamble, de hecho el método produce una única estimación por iteración. Sin embargo, las repeticiones nos dan una idea de la incerteza de las estimaciones la cual se estabiliza alrededor de los 500 ciclos. Los tres métodos estiman correctamente el valor real del parámetro aunque los basados en importance sampling muestran una ligera subestimación. La velocidad de convergencia y la variabilidad en las estimaciones es también similar en todos los métodos.

Efecto de la tasa de aprendizaje

Para evaluar el efecto de la tasa de aprendizaje en las estimaciones realizamos un experimento similar al anterior: estimamos una matriz **Q** diagonal considerando **R** conocida en el modelo Lorenz-63. En cada simulación utilizamos el mismo set de observaciones sintéticas pero cambiamos el valor de α de la tasa de aprendizaje. En la Figura 3.8 se muestran las estimaciones para la media de la diagonal de **Q**. Para valores mayores de α la convergencia es lenta e incluso puede no lograrse en 10000 ciclos de asimilación. Por otro lado, para valores más pequeños se obtiene una convergencia mucho más rápida. Las estimaciones que produce el método son una ponderación entre la estimación de la iteración anterior con el promedio de los estadísticos suficientes correspondientes a la observación que está siendo procesada. Cuando α es más cercano a 1 se da más peso a la estimación anterior y menos a la información introducida por la última observación. Esto puede interpretarse como que el método tiene más memoria por lo que las nuevas iteraciones no cambian sustancialmente las estimaciones, con lo cual estas resultan menos ruidosas pero la convergencia se realentiza. Por el contrario, para valores menores de α se le da mayor peso a los estadísticos suficientes correspondientes a la última observación y menos a las estimaciones anteriores. Esto acelera la convergencia pero las trayectorias se hacen más ruidosas debido al error de muestreo.

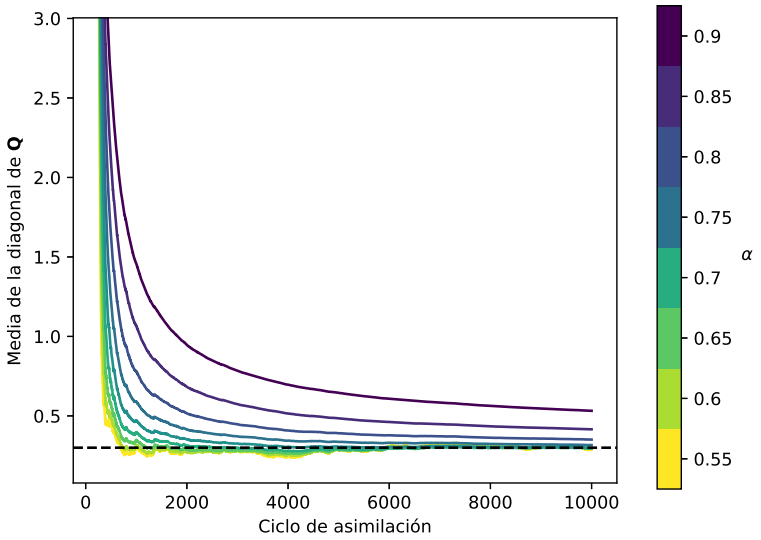

Figura 3.8: Estimaciones de la media de la diagonal de **Q** con OSS-EnKF (panel superior), IS-EnKF (panel medio) y IS-VMPF (panel inferior).

Estimación de covarianzas

En aplicaciones geofísicas es normal que las variables estén correlacionadas espacialmente. En estos sistemas es natural que el error de modelo también reproduzca estas correlaciones. Esto resulta en que la matriz **Q** tenga una estructura con valores no nulos fuera de la diagonal. Para este conjunto de experimentos consideraremos el modelo Lorenz-96 (Lorenz, 1996), determinado por las siguientes ecuaciones diferenciales:

$$\frac{dX_n}{dt} = X_{n-1}(X_{n+1} - X_{n-2}) - X_n + F \quad \forall n = 1, ..., N_x \tag{3.23}$$

Tomaremos condiciones de frontera periódicas, es decir, $X_{-1} = X_{N_x-1}, X_0 = X_{N_x}, X_1 = X_{N_x+1}$. Este modelo busca emular el comportamiento de una variable atmosférica en un círculo de latitud con un forzado externo representado por F. Consideraremos, además de la varianza del error, representado por valores positivos en la diagonal de **Q**, covarianzas positivas e uniformes entre variables adyacentes. Esto significa que $\mathbf{Q}_{i,(i+1)\%N_x} = \sigma^2_{ady} > 0 \quad \forall i = 1, ..., N_x$. Los valores en la diagonal también serán uniformes de manera que $\mathbf{Q}_{i,i} = \sigma^2_{diag} > 0 \quad \forall i = 1, ..., N_x$. Para estimar esta matriz utilizamos las tres implementaciones del EM *online* y también con 25 iteraciones del EM *offline* en la implementación del Algoritmo 6. Notemos que la versión *online* es una aproximación del *offline* por lo que la segunda constituye una base de comparación razonable para la primera.

A pesar de que la matriz **Q** tiene una estructura que admite una representación de sólo dos parámetros, las estimaciones producidas por cualquiera de los métodos son entrada por entrada (salvo simetría pues los estadísticos suficientes son simétricos). Para verificar que el comportamiento de las estimaciones en cada entrada es consistente mostramos en la Figura 3.9 las trayectorias correspondientes. Podemos observar que estas se agrupan correctamente alrededor de los valores de la matriz **Q** original. En la Figura 3.10 se puede ver la distancia Frobenious de las estimaciones de todos los métodos. Los errores de las estimaciones en los métodos secuenciales

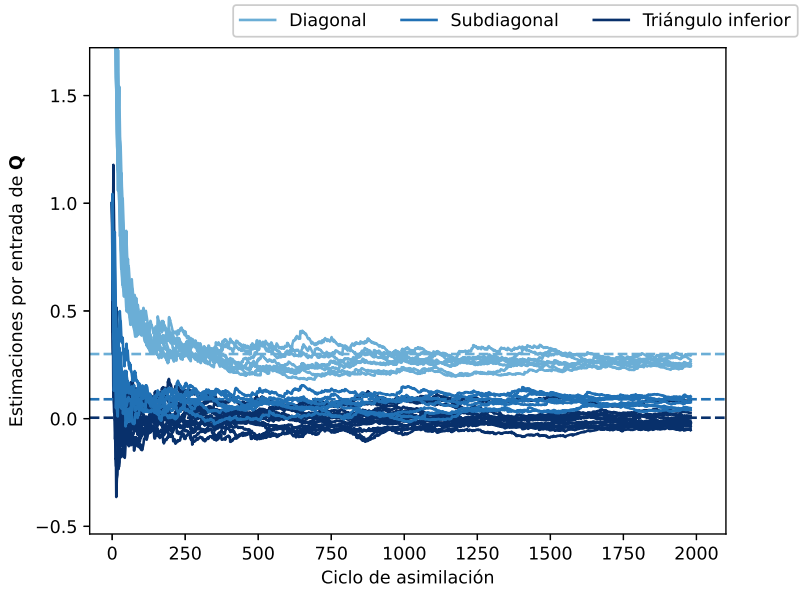

Figura 3.9: Estimaciones entrada por entrada de **Q** en líneas continuas diferenciados por tono de acuerdo al sector de la matriz al que pertenecen. Las líneas intermitentes indican los valores reales correspondientes para cada sector de la matriz.

se aproximan con velocidad similar entre ellos al error obtenido por el EM *offline*. Además, para cada estimación de **Q** se realizó un filtrado con el EnKF con **Q** fijada en el valor de la estimación. Del resultado de estos filtrados computamos el RMSE de las variables de estado y la log-verosimilitud obtenida. Se puede ver en la Figura 3.10 que la performance según estas métricas, son similares a las del EM *batch*.

Finalmente, realizamos un experimento similar pero considerando que la matriz **Q** varía lentamente en el tiempo de acuerdo a una función sigmoidea. La versión *batch* del EM produce una única estimación del parámetro para toda la ventana observacional por lo que no sería adecuado usarlo en este escenario. Por el otro lado los métodos secuenciales, al producir una estimación por cada observación son capaces de capturar cambios en el parámetro. Notemos sin embargo que las estimaciones, al tener memoria no pueden capturar cambios muy abruptos. En la Figura 3.11 se observa que todos los métodos responden a los cambios en **Q** aunque, especialmente para los valores en la diagonal se produce una subestimación respecto al valor real. Esto podría deberse a que el efecto de la memoria del método provoque que las estimaciones aún tengan una tendencia hacia los valores más tempranos y más pequeños de la ventana temporal.

Estimación conjunta de Q y R

Hasta ahora nos hemos enfocado en la especificación del error de modelo puesto que es una fuente de incerteza de la cual, en la mayor parte de los casos, no tenemos forma de determinar, debido a que suele provenir del desconocimiento del proceso subyacente. El error observacional por otro lado, muchas veces se puede aproximar pues conocemos mejor el proceso observacional. Sin embargo, como hemos visto en la Sección 3.1 la especificación de ambas incertezas es importante y de hecho existe interacción entre ambas cantidades. Por esto se hace importante estudiar la estimación conjunta de **Q** y **R**. En este experimento lo hacemos tomando a ambas como

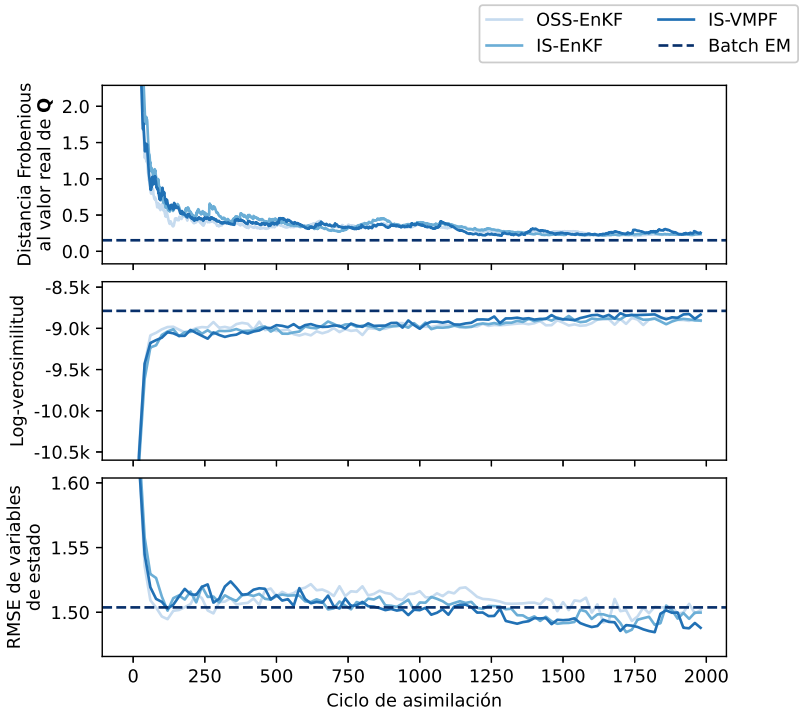

Figura 3.10: Panel superior: distancia Frobenious al verdadero valor de $\mathbf{Q}$. Panel medio: Log-verosimilitud que se obtendría si se usara la estimación en cada ciclo para filtrar toda la ventana. Panel inferior: similar al panel medio pero considerando RMSE.

matrices diagonales múltiplos de la identidad: $\mathbf{Q} = \sigma^2_{\mathbf{Q}}\mathbf{I}$ y $\mathbf{R} = \sigma^2_{\mathbf{R}}\mathbf{I}$. El modelo utilizado es el Lorenz-96 con 8 variables, todas observadas. Consideramos dos escenarios: en el primero tenemos errores observacionales y de modelo similares, $\sigma^2_{\mathbf{Q}} = 0.3$ y $\sigma^2_{\mathbf{R}} = 0.5$; mientras que en el segundo tomamos un error observacional mayor, $\sigma^2_{\mathbf{R}} = 1.5$. El efecto que se encuentra cuando $\mathbf{Q}$ y $\mathbf{R}$ son similares es que la verosimilitud tiene un máximo mejor definido. En la Figura 3.12 vemos los contornos de la verosimilitud en función de $\sigma^2_{\mathbf{Q}}$ y $\sigma^2_{\mathbf{R}}$ para ambos casos. El panel izquierdo corresponde al escenario de errores similares y se puede observar que el máximo es más aguzado que el escenario con errores disímiles del panel derecho. Además graficamos las trayectorias de las estimaciones para distintos valores iniciales. Se puede ver que cuando el máximo está mejor definido las estimaciones son más precisas. Finalmente notamos que los valores de la verosimilitud son menores en el caso de mayor error observacional, lo cual es esperable porque las observaciones son menos precisas.

3.3.4 Discusión

La metodología de EM *online* que desarrollamos muestra resultados prometedores y da respuesta a algunos de los desafíos que típicamente se presentan en escenarios de asimilación de datos. La técnica permite evitar el uso de suavizadores sobre grandes ventanas temporales lo que la hace computacionalmente menos costosa que las versiones por lotes del EM y especialmente adecuada para contextos de asimilación de datos secuencial. Es importante notar que la performance puede depender de la tasa de aprendizaje utilizada. La técnica permite la estimación de varianzas pero también covarianzas en los errores de las variables de estado y adicionalmente

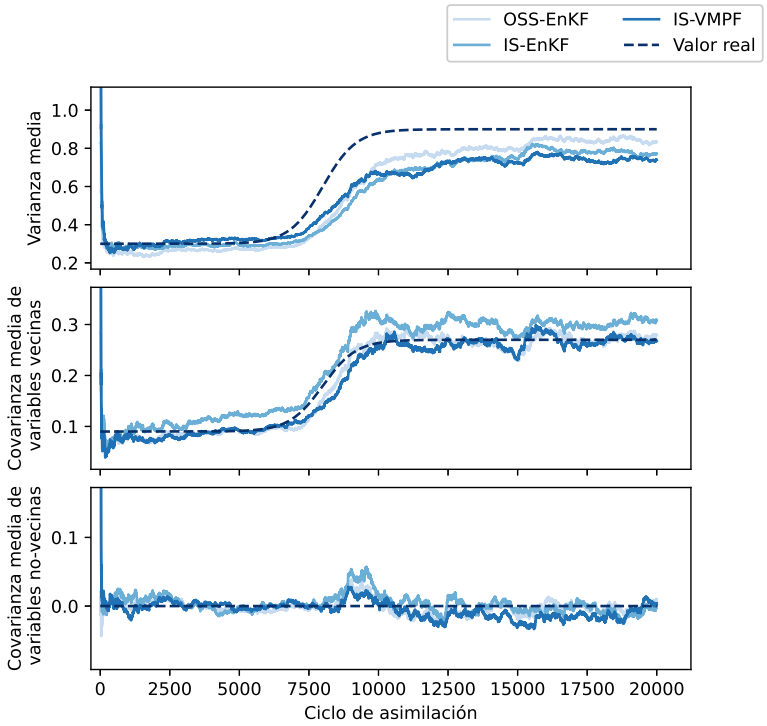

Figura 3.11: Estimaciones para las distintas implementaciones de una matriz **Q** que varía en el tiempo y con covarianzas fuera de la diagonal. En el panel superior se muestran las estimaciones correspondientes a la diagonal, en el medio de las covarianzas que corresponden a variables vecinas y en el inferior las del resto de la matriz (covarianzas de variables con más de un grado de separación).

detecta cambios temporales lentos en los parámetros estocásticos. No evaluamos el desempeño del método en espacios de alta dimensionalidad ya que en estos casos las mismas técnicas de asimilación pueden comenzar a tener peor rendimiento debido a que se dificulta la representación de distribuciones mediante muestras en este tipo de espacios.

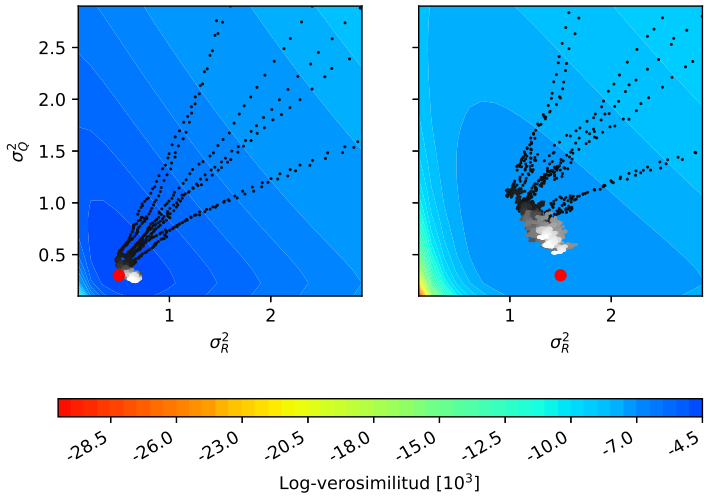

Figura 3.12: Estimaciones correspondientes a $\sigma^2_{\mathbf{Q}}$ y $\sigma^2_{\mathbf{R}}$ sobre contornos de la verosimilitud para los dos escenarios planteados. El panel derecho corresponde a los valores reales $\sigma^2_{\mathbf{Q}} = 0.3$ y $\sigma^2_{\mathbf{R}} = 0.5$ y el izquierdo a $\sigma^2_{\mathbf{Q}} = 0.3$ y $\sigma^2_{\mathbf{R}} = 1.5$ Las estimaciones están indicadas con puntos en escala de grises: comienzan con tonos oscuros y terminan en tonos claros. La barra de colores indica los valores de la verosimilitud dividida por un factor de 10^3. Los puntos rojos indican los valores reales.

Capítulo 4

Modelos epidemiológicos

4.1 Modelos compartimentales

Para modelar la propagación de enfermedades infecciosas, es muy habitual utilizar modelos compartimentales basados en ecuaciones diferenciales. Estos pueden ser utilizados para predecir y comprender situaciones epidemiológicas, ayudar a la toma de decisiones y simular escenarios hipotéticos así como también se pueden utilizar para estimar parámetros. Su origen se remonta a los años 20 del siglo pasado y suele atribuirse a Kermack et al., 1927. Estos consisten en modelar la población distinguiendo subpoblaciones de acuerdo a categorías epidemiológicas. Por ejemplo, el emblemático modelo SIR, distingue a las subpoblaciones susceptible (S), infectada (I) y recuperada (R). La evolución de estas variables se puede expresar mediante el siguiente sistema de ecuaciones diferenciales,

$$\frac{\partial S}{\partial t} = -\beta \frac{SI}{N} \tag{4.1}$$

$$\frac{\partial I}{\partial t} = \beta \frac{SI}{N} - \gamma I \tag{4.2}$$

$$\frac{\partial R}{\partial t} = \gamma I \tag{4.3}$$

Estas ecuaciones suponen una población constante de tamaño N: notemos que las ecuaciones suman 0 por lo que la población total se mantiene. La velocidad con que los individuos susceptibles se infectan es proporcional a la cantidad total de susceptibles, S, y a la proporción de infectados de la población, $\frac{I}{N}$. La constante de proporcionalidad β es llamada la tasa de infección y cuantifica la transmisibilidad de la enfermedad. Puede ser pensada como la cantidad de infecciones que produce un infectado por unidad de tiempo. Por su parte, la velocidad con que los infectados se recuperan es proporcional a la cantidad de infectados y la constante de proporcionalidad γ es llamada tasa de recuperación. Esto tiene como consecuencia que el tiempo de permanencia esperado en la clase I sea $\frac{1}{\gamma}$. El sistema resulta en una epidemia cuando $\frac{\partial I}{\partial t} > 0$ y esto sucede cuando $\frac{\beta S(0)}{\gamma} > N$. El número reproductivo básico, R_0 en un modelo epidemiológico suele definirse como la cantidad de infectados directos que provocaría un infectado en una población totalmente susceptible. En este caso se puede computar de manera explícita como $R_0 = \frac{\beta}{\gamma}$. Esto se puede pensar como que, por cada unidad de tiempo en el intervalo en que el infectado está en el compartimento I, se infecta de manera directa a β susceptibles. El umbral en el comportamiento del sistema depende entonces de este valor pues para que se de un pico epidémico se debe dar que $R_0 S(0) > N$. En general $t = 0$ representa un momento en el que la enfermedad está surgiendo y por lo tanto $S(0) \approx N$. Debido a esto se suele considerar que la condición para tener una epidemia es cuando $R_0 > 1$.

El modelo SIR, aunque sencillo, merece mención pues tiene los elementos básicos con los que se puede construir un modelo compartimental. Como vimos, la entrada y salida de cada compartimento se describe con una ecuación diferencial y con esta idea se pueden diseñar distintas estructuras de subpoblaciones para describir las situaciones que se deseen modelar. Es común ver en este tipo de modelos la subpoblación de los decesos D que permite distinguir a la porción de la población que muere a causa de la enfermedad, o la subpoblación de los expuestos E se suele ver en modelos para enfermedades virales en que las personas se infectan e incuban el virus sin ser infecciosas antes de pasar al compartimento I. De esta manera se habilitan configuraciones como el SEIRD que considera estas subpoblaciones, o el SIS que no tiene en cuenta una categoría de recuperados sino que considera que los individuos, al curarse de la enfermedad vuelven a ser susceptibles. Otros consideran compartimentos que corresponden a los inmunizados mediante vacunación. Hay una gran diversidad de modelos de compartimentales de esta naturaleza con distintas elecciones para las categorías epidemiológicas y para las dinámicas de transpaso de uno compartimento a otro: en (Murray, 2007) se pueden encontrar numerosas configuraciones posibles de este concepto para distntos escenarios. Habitualmente se representa con un diagrama de flujo cuales son los compartimientos y qué direcciones toma el cambio de estado epidemiológico de acuerdo a las características de la enfermedad. Por ejemplo, en la Figura 4.1 podemos ver esta representación para un modelo SIR.

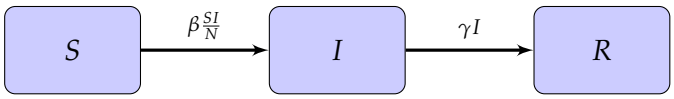

Figura 4.1: Diagrama para un modelo SIR

En, general podemos distinguir 3 tipos distintos de compartimentos: aquellos que actúan de fuente, es decir que solo decrecen o se mantienen, y por lo tanto los términos en las ecuaciones diferenciales son sólo negativos, los compartimentos intermedios, cuyas ecuaciones incluyen términos positivos y negativos y finalmente los compartimentos que actúan de sumidero, es decir que sólo crecen o se mantienen y sólo incluyen términos positivos en sus ecuaciones. Por ejemplo, en el modelo SIR el compartimento S actúa a modo de fuente, I es un compartimento intermedio y R actúa de sumidero. Típicamente las trayectorias de las fuentes tienen una forma sigmoidea decreciente y las de los sumideros son sigmoideas crecientes mientras que los compartimientos intermedios suelen exhibir un comportamiento con forma de campana. En la Figura 4.2 se puede observar esto en las trayectorias generadas por un modelo SIR.

La incorporación de compartimentos dota de adaptabilidad a estos modelos. Sin embargo, se considera que las subpoblaciones representadas en cada uno de estos compartimentos se mezcla de manera homogénea con lo cual no se representan los efectos que provienen de las complejidades del comportamiento de los individuos. Como veremos más adelante, los modelos basados en agentes se focalizan en modelar cada individuo de manera de dar cuenta de las consecuencias de las interacciones en esta micro-escala. Por otro lado, la representación mediante ecuaciones da la posibilidad de analizar formalmente el comportamiento del sistema para, por ejemplo, identificar puntos de equilibrio, umbrales entre distintos tipos de soluciones o cálculos explícitos del número reproductivo básico (Murray, 2007).

Además de la incorporación de subpoblaciones relacionadas a los estados epidemiológicos, los modelos compartimentales pueden incluir otras características.

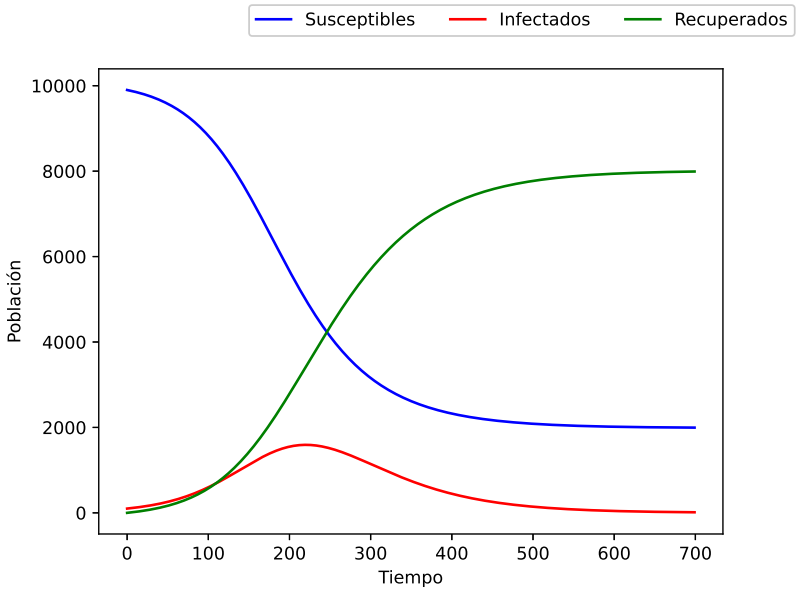

Figura 4.2: Trayectorias de susceptibles, infectados y recuperados para un modelo SIR básico

Por ejemplo, se pueden utilizar variables espaciales acopladas a términos de difusión en las ecuaciones para representar la dispersión territorial de las enfermedades (Noble, 1974). También se utiliza estratificación por edades para diferenciar el efecto de la enfermedad en distintos sectores de la población y codificar el contacto entre los distintos rangos etarios (Hethcote, 2000). Además, se puede modelar la espacialidad a través de la determinación de áreas, por ejemplo barrios en una ciudad, cada una con sus propias subpoblaciones de susceptibles, infectados etc. y describiendo la interacción entre estos lugares (Klepac et al., 2018). Con el advenimiento de la pandemia de COVID-19 se desarrollaron un gran número de modelos compartimentales con características especiales para expresar determinados rasgos propios de esta enfermedad y que se han utilizado para diversos fines tales como la estimación de parámetros epidemiológicos, la evaluación de medidas de control o la predicción de tendencias en los picos de contagios. Por ejemplo, el modelo de Evensen et al., 2020 utiliza estratificación por edades y compartimentos para representar subpoblaciones bajo distintos tipos de cuarentena y es utilizado para estimar R_0 y hacer predicciones en distintos escenarios en diferentes países. El modelo de Arenas et al., 2020 también incluye estratificación por edades pero también incorpora patrones de movimiento en distintas regiones de España, compartimentos para los hospitalizados y otras características incluidas para adaptar el esquema al COVID-19. En general, cuando tenemos una segmentación de la población en K grupos, ya sea rango etario, locación o algún otro, tenemos una réplica del sistema de ecuaciones por cada uno de ellos. La interacción entre estos se puede representar mediante una matriz $K \times K$. Si extendemos las ecuaciones del modelo SIR a un modelo multi-grupos tendremos las

siguientes ecuaciones para $i = 1, ..., K$:

$$\frac{\partial S_i}{\partial t} = -\beta \sum_{j=1}^{K} c_{ij} \frac{S_i I_j}{N} \tag{4.4}$$

$$\frac{\partial I_i}{\partial t} = \beta \sum_{j=1}^{K} c_{ij} \frac{S_i I_j}{N} - \gamma I_i \tag{4.5}$$

$$\frac{\partial R_i}{\partial t} = \gamma I_i \tag{4.6}$$

donde c es la matriz de conectividad, de manera que la interacción entre el grupo i y y el j está cuantificado por c_{ij}.

4.2 Inferencia en modelos compartimentales

Los modelos epidemiológicos por sí solos pueden ser útiles para entender algunos fenómenos respecto a las dinámicas de contagio o los posibles efectos de medidas de control en escenarios hipotéticos. Sin embargo, la parametrización que se utilice puede cambiar drásticamente el comportamiento de los modelos y se hace importante la estimación de parámetros para un mejor ajuste del modelo a los datos o para obtener pronósticos más certeros. Por estos motivos es normal acoplar al modelo alguna técnica de inferencia que permita realizar esta tarea. En particular, la asimilación de datos provee, por un lado técnicas para estimar parámetros como el estado aumentado y por otro mejora los pronósticos de las mismas variables de estado incorporando la información de las observaciones. Además de esto, la asimilación de datos está diseñada de manera que es natural el tratamiento de variables no observadas, y en un sistema epidemiológico es normal que haya observaciones incompletas (por ejemplo que sólo se observe la subpoblación de infectados y muertos pero que no se observe ninguna de las otras).

Los trabajos Ionides et al., 2006 y posteriormente Shaman y Karspeck, 2012; Shaman, Karspeck et al., 2013 constituyen excelentes ejemplos de la aplicación de técnicas modernas de asimilación de datos en modelos epidemiológicos. En Ionides et al., 2006 se utiliza, para un modelo compartimental para el cólera, la técnica de filtros iterados para estimar variables de estado y parámetros estocásticos y físicos. La técnica está basada en el uso de estado aumentado pero con la variante de que se pasa repetidas veces el filtro y en cada iteración se reduce la varianza del parámetro de estado aumentado de tal manera que el valor de la estimación se va estabilizando en un valor fijo que corresponde al máximo de la verosimilitud. Además se utiliza un filtro de partículas lo cual permite la inferencia en escenarios de no linealidad. Por otro lado, en los trabajos de Shaman y Karspeck, 2012; Shaman, Karspeck et al., 2013 se aplica el EAKF, una variante del EnKF, a un modelo SIRS para predecir picos de gripe y estimar parámetros relacionados a R_0 mediante estado aumentado. Más recientemente, ha habido aplicaciones para modelos de COVID-19. En Li, Pei et al., 2020 se utilizan filtros iterados pero con el EAKF para estimar casos no reportados en China. En Evensen et al., 2020 se aplica una técnica iterativa basada en suavizadores de Kalman por ensambles llamada ESMDA (*Ensemble Smoother with Multiple Data Assimilation*) para calibrar la parametrización del modelo y, en particular, estimar el número de reproducción efectivo en un modelo compartimental. Por otro lado en Ghostine et al., 2021 se utiliza el EnKF para estudiar el efecto de la vacunación con datos de Arabia Saudita.

A pesar de que las aplicaciones de técnicas de asimilación de datos a modelos epidemiológicos se comienzan a popularizar, no se ha hecho particular foco en la estimación de errores observacionales y de modelo. Estas cantidades no sólo tienen interés en sí mismo para la comprensión de las incertezas del sistema sino que, como hemos visto en el Capítulo 3, también afectan en la performance de la asimilación. En particular, el error observacional en modelos epidemiológicos tiene la complejidad de estar relacionado al reporte de casos lo cual puede ser una fuente de información poco precisa y cuya incerteza varía en el tiempo. Por estas cualidades realizamos para esta tesis experimentos preliminares de aplicación de el algoritmo EM online presentado en la Sección 3.3.2 cuyos resultados presentamos a continuación.

4.2.1 Experimento: observaciones sintéticas

Para mostrar algunas de las capacidades de el EnKF acoplado con el EM online en un modelo epidemiológico realizamos un experimento con observaciones sintéticas sobre un modelo SEIRD definido por las siguientes ecuaciones diferenciales:

$$\frac{\partial S}{\partial t} = -\beta \frac{SI}{N} \tag{4.7}$$
$$\frac{\partial E}{\partial t} = \beta \frac{SI}{N} - \gamma_E E \tag{4.8}$$
$$\frac{\partial I}{\partial t} = \gamma_E E - \gamma_I I \tag{4.9}$$
$$\frac{\partial R}{\partial t} = \alpha \gamma_I I \tag{4.10}$$
$$\frac{\partial D}{\partial t} = (1-\alpha)\gamma_I I \tag{4.11}$$

donde S, E, I, R y D representan a los susceptibles, expuestos, infectados, recuperados y muertos respectivamente. β es la tasa de infección, γ_E determina la velocidad con que los expuestos pasan a ser infectados y γ_I la tasa a la que los infectados se recuperan o mueren. La proporción de infectados que mueren es α y la que se recupera $1-\alpha$. El sistema se puede representar mediante el diagrama de la Figura 4.3.

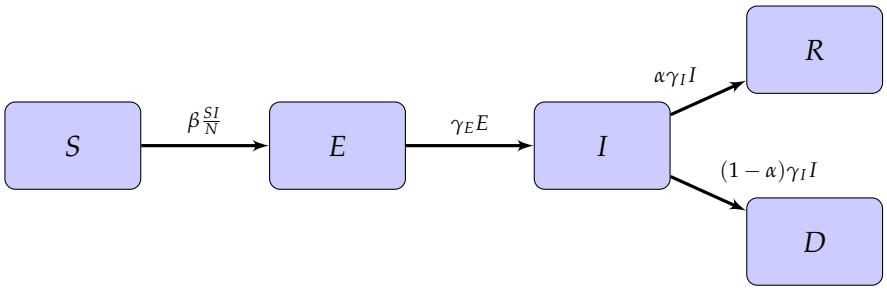

Figura 4.3: Diagrama para un modelo SEIRD

Consideramos que las observaciones corresponden a los infectados acumulados y a los muertos. Los infectados acumulados, que llamaremos I_{tot} pueden ser computados como la suma de las variables I, R y D. Por lo tanto el operador observacional se puede escribir en forma matricial como:

$$\mathbf{H} = \begin{pmatrix} 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Al error observacional lo consideraremos Gaussiano y aditivo mientras que a $\mathbf{R}_t$, la matriz que especifica su varianza a tiempo t, la consideraremos diagonal, es decir sin correlación entre los errores. La varianza para el error en la primera variable observada la tomaremos proporcional a los infectados confirmados diarios, i.e., $(\mathbf{R}_t)_{1,1} \propto I_{tot}(t) - I_{tot}(t-1)$. Análogamente, para la segunda variable observada tomamos $(\mathbf{R}_t)_{2,2} \propto D(t) - D(t-1)$. El parámetro β será definido como una función constante a trozos. En particular, tendrá un valor de 0.35 en toda la ventana temporal excepto por el intervalo donde $t \in (400, 600)$ donde su valor será 0.3. Esto tiene el efecto de que el número de casos aumente, después amaine temporalmente para luego volver a crecer. Esto puede ser utilizado para representar de manera sencilla un escenario de confinamiento estricto y posterior relajación de la medida. Esta configuración para β y $\mathbf{R}_t$ se utilizará para generar las observaciones sintéticas pero será desconocida para el sistema de asimilación: β se estimará mediante estado aumentado mientras que $\mathbf{R}_t$ a través del EM *online*.

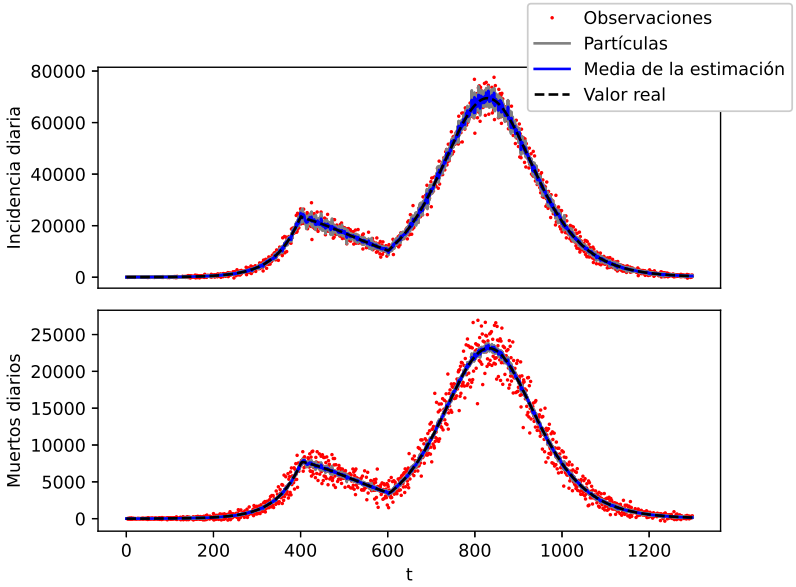

Figura 4.4: Trayectorias reales, observaciones y estimaciones para los infectados diarios (incidencia diaria) y muertos diarios utilizando el modelo SEIRD

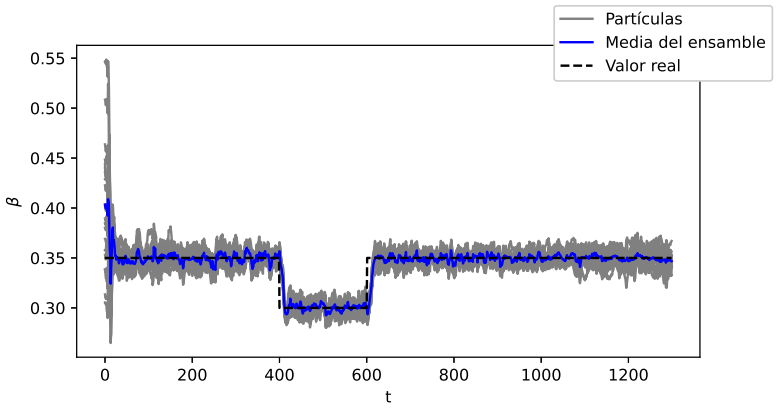

Figura 4.5: Valor real y estimaciones de la tasa de infección β

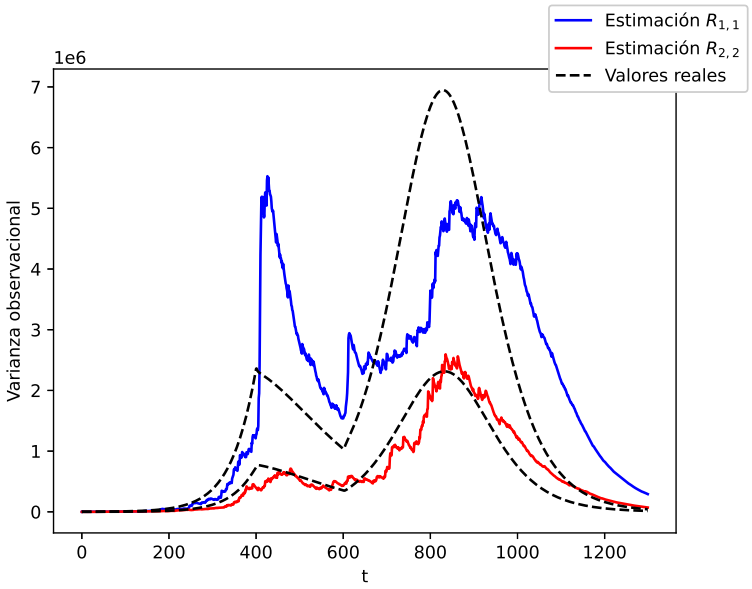

Figura 4.6: Valores reales y estimados mediante EM online de las varianzas de errores observacionales. Estas corresponden a los dos valores de la diagonal de las matrices de covarianzas $\mathbf{R}_t$

En la Figura 4.4 se muestran los valores observados y reales de los infectados y muertos diarios junto sus estimaciones. Se ve que el error observacional es proporcional al valor real como se especificó anteriormente y que las estimaciones de las variables sincronizan muy bien con los valores reales. El efecto de variación temporal sobre la tasa de infección β explica el efecto de que el sistema produzca dos picos epidémicos. En la Figura 4.5 se puede ver que las estimaciones de β pueden capturar correctamente el cambio temporal del parámetro. Existe un retardo en las estimaciones para sincronizar luego de los cambios abruptos de la tasa de infección lo cual es habitual para técnicas de estado aumentado debido a que la asimilación estima los parámetros a través de sus correlaciones con las variables observadas por lo que tiene cierta inercia hacia valores previos y esto se corrige a medida que las observaciones se van incorporando. Finalmente, en la Figura 4.6 tenemos las estimaciones de $\mathbf{R}_t$ producidas por el EM *online* junto a los valores reales utilizados para generar las observaciones. A pesar de haber desvíos sustanciales de las estimaciones respecto a las variaciones temporales reales de la varianza es evidente que el algoritmo EM percibe los cambios en $\mathbf{R}_t$ y da aproximaciones que están en escalas similares. En este experimento utilizamos una tasa de aprendizaje $\alpha = 0.6$ lo que configura al algoritmo para tener poca memoria, dándole más importancia a las observaciones que están siendo procesadas y no tanto a las estimaciones anteriores. Realizamos réplicas de este mismo experimento con otros valores de α y obtuvimos estimaciones menos ruidosas pero que reaccionan más lentamente a los cambios del parámetro. La inercia observada es mayor al caso de los parametros deterministicos ya que a la inercia propia de la asimilación secuencial de datos, se le suma el retardo que incorpora la memoria del EM *online*.

4.2.2 Experimento: datos COVID-19 de Argentina

Como demostración de la potencial viabilidad de la técnica en modelos epidemiológicos realizamos otro experimento similar al de la Sección 4.2.1 pero utilizando

los datos de COVID-19 en Argentina provistos por el Ministerio de Salud. El experimento tiene una configuración similar al de observaciones sintéticas que describimos anteriormente pero además incorporamos al estado aumentado a la tasa de mortalidad que llamaremos γ_D y es computable como $(1-\alpha)\gamma_I I$ según las Ecuaciones 4.7.

Cuando se tienen parámetros en el estado aumentado (tal como se expone en la Sección 3.2) las caminatas aleatorias que se utilizan para que las estimaciones exploren el espacio paramétrico tienen el efecto de actuar como error de modelo. Por lo tanto, estas interactúan con el resto del sistema, con las estimaciones de $\mathbf{R}$, y con el comportamiento de la media y dispersión de las variables de estado (ver Sección 3.1). En este experimento parametrizamos las dos caminatas aleatorias con un único valor σ^2. La amplitud de los pasos es Gaussiana con media 0 y varianza σ^2 para β y $\sigma^2/100$ para γ_D. Para estudiar su efecto corrimos el experimento para distintos valores de σ^2.

En la Figura 4.7 se grafican los infectados y muertos diarios observados y estimados para el caso de $\sigma^2 = 0.015$. Las trayectorias estimadas están prácticamente superpuestas a las observaciones y la escala impide apreciar la dispersión del ensamble sin embargo este no está colapsado. Los resultados para otros valores de σ^2 son similares pero la dispersión del ensamble tiende a ser menor cuando se aumenta su valor. Esto se debe a que σ actúa como error de modelo y para valores mayores el ensamble tiende sobreestimar la verosimilitud de las observaciones.

Las estimaciones de β y γ_D están en la Figura 4.8. El comportamiento global de los parámetros es similar para todos los valores de σ^2 pero a medida que este valor aumenta las estimaciones son más ruidosas y con mayor varianza. El aumento de la varianza se debe a que σ cuantifica la dispersión del ensamble del pronóstico. El ruido en las estimaciones se origina en que, al estar en un escenario de sobreajuste, el parámetro absorbe el ruido observacional. La tasa de infección es mayor al comienzo de la pandemia y alrededor de los picos de casos, notablemente en el que corresponde a finales de 2021. Por otro lado la tasa de mortalidad es mayor al comienzo de la ventana temporal y se va reduciendo en el tiempo. Esto significa que el incremento en muertos en la última ola se explicaría, según este modelo, por el aumento de casos pero no por una mayor letalidad de la enfermedad.

Finalmente, en la Figura 4.9 se pueden ver las estimaciones de las varianzas de los errores observacionales $(\mathbf{R}_t)_{1,1}$ y $(\mathbf{R}_t)_{2,2}$. En este caso es notable como el valor de σ^2 afecta a las estimaciones que produce el EM *online*: el comportamiento es muy similar para distintos valores de σ^2 pero la escala es muy diferente. Esto se debe a la interacción entre el error de modelo y observacional. Es esperable que mientras mayor sea el error de modelo el sistema le de más importancia a las observaciones. El error que corresponde a la primera variable observada, los infectados acumulados, tiene un pico notable alrededor de la primera ola y otro en la ola de finales de 2021 pero este último se hace menos importante a medida que aumenta σ^2. Por su parte, el error que corresponde a los fallecidos acumulados es mayor al comienzo de la pandemia y luego se reduce y se estabiliza.

Los resultados dependen de la elección de las caminatas aleatorias en los parámetros dentro del estado aumentado ya que estos tienen el efecto de inyectar error de modelo al sistema. De hecho, podemos ver que al aumentar el valor de σ^2 se pierde credibilidad en el modelo porque este pasa a tener mayor error. Esto resulta en que el sistema de asimilación de datos asigne mayor importancia a las observaciones y que se reduzca el error observacional estimado por el EM *online*. En la Figura 4.8 se aprecia que, al aumentar la varianza de las caminatas aleatorias, las estimaciones de

los parámetros se hacen más ruidosas: esto es un comportamiento típico de sobreajuste a las observaciones. El sistema le da demasiada credibilidad a los datos y por lo tanto las trayectorias de las variables de estado (incluidas las del estado aumentado) absorben el error de las observaciones.

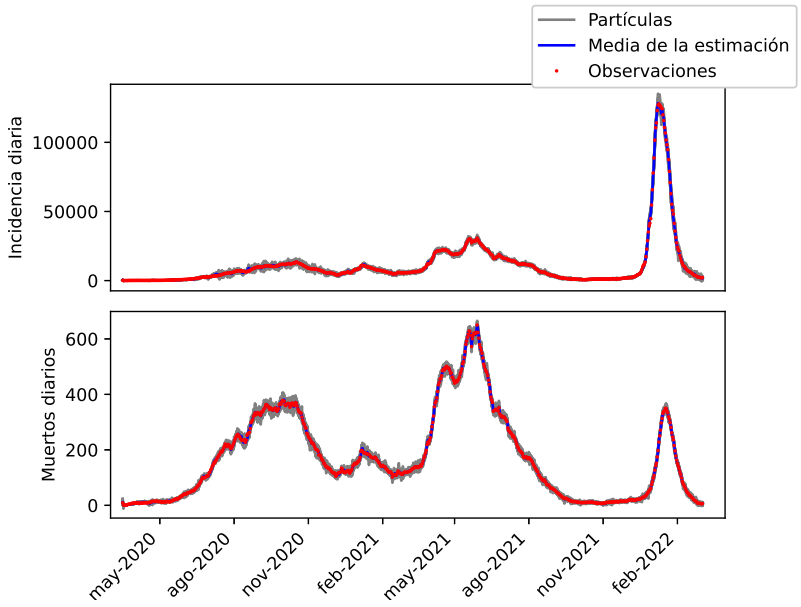

Figura 4.7: Observaciones y estimaciones para los infectados diarios (incidencia diaria) y muertos diarios utilizando el modelo SEIRD sobre los datos de COVID-19 en Argentina

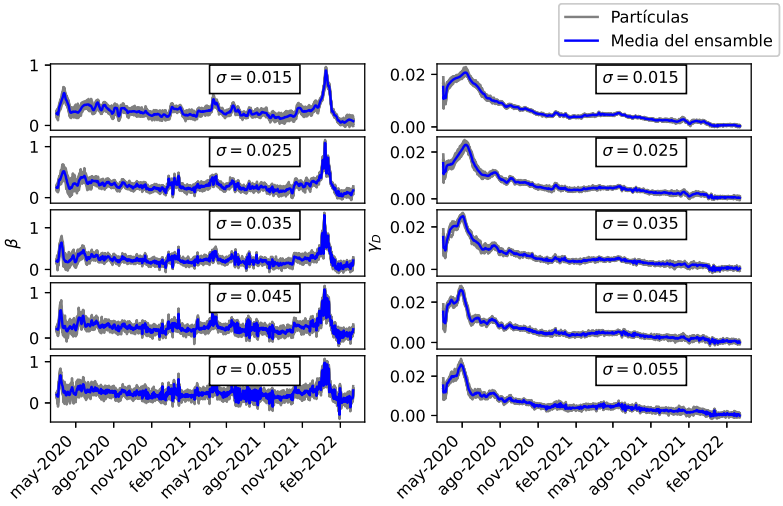

Figura 4.8: Estimaciones de la tasa de infección β y la tasa de fatalidad γ_D

4.2.3 Discusión

Diseñamos estos experimentos para mostrar la potencial viabilidad del EM *online* para estimación de errores en sistemas de asimilación de datos sobre modelos epidemiológicos. Este aspecto es habitualmente pasado por alto en trabajos de inferencia en este tipo de sistemas. En el experimento con observaciones sintéticas, salvo por la parametrización que se calibra con estado aumentado, el modelo utilizado

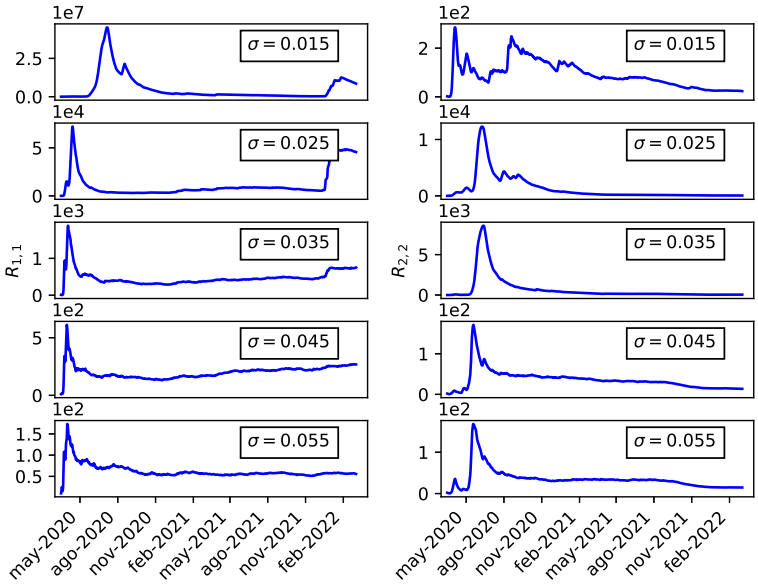

Figura 4.9: Estimaciones del EM online de las varianzas de errores observacionales. Estas corresponden a los dos valores de la diagonal de las matrices de covarianzas $\mathbf{R}_t$

es el mismo que se usa para generar los datos. En el caso de tener datos reales no es posible especificar un modelo perfecto por lo que los parámetros en el estado aumentado cumplen el rol de calibrar otras posibles deficiencias del modelo. Los resultados obtenidos evidencian la interdependencia entre el error observacional y de modelo y cómo estos afectan la performance general de las estimaciones. Una posibilidad es incorporar error de modelo aditivo y estimar su matriz de covarianza $\mathbf{Q}_t$ con el EM *online* y de hecho algunos experimentos preliminares sugieren que esto es posible. Además, las varianzas de las caminatas aleatorias en el estado aumentado podrían potencialmente ser estimadas de manera adaptativa como parte de la matriz $\mathbf{Q}_t$. Otra posibilidad más sencilla para dar cuenta del error de modelo es la utilización de un factor de inflación en el ensamble (ver Sección 2.4.2).

Capítulo 5

Asimilación de datos en modelos basados en agentes

5.1 Modelos basados en agentes

A diferencia de los modelos de predicción con ecuaciones diferenciales, los modelos basados en agentes (conocidos como ABMs por sus siglas en inglés) se basan en un paradigma diferente. Estos modelan de manera explícita las características y comportamiento de individuos o agentes que interactúan. El comportamiento conjunto de la población de agentes se utiliza como modelo de un sistema complejo (Bonabeau, 2002). Incluso reglas de interacción simples pueden resultar en que los agentes se organicen de manera autónoma y que el comportamiento colectivo de la población emerja de manera *bottom-up* desde la escala micro del modelado de los individuos (Helbing, 2012). En general, los ABMs requieren una gran capacidad computacional ya que las poblaciones de agentes pueden ser muy grandes. Actualmente es factible utilizarlos de manera operacional y existen ejemplos de su aplicación en epidemiología, ecología, economía y sociología (Epstein y Axtell, 1996; Grimm et al., 2005; Tesfatsion y Judd, 2006; Vynnycky y White, 2010).

Los modelos de agentes tienen dos componentes principales que los describen:

1. Las descripciones de los agentes mediante atributos
2. Las reglas de comportamiento e interacción

Los agentes pueden ser vistos como un tipo de dato con diferentes campos de manera que cada individuo se define por el valor de estos atributos. Para completar al modelo se agrega el comportamiento e interacción de los agentes, en general mediante reglas que pueden contener componentes estocásticos. Las acciones que realicen los individuos bajo estas disposiciones provocarán cambios en los valores de sus atributos. En principio estos atributos pueden ser cualquier tipo de dato, de manera que los agentes pueden ser programados como tipos de datos compuestos (como estructuras en C) o incluso a través de clases en lenguajes orientados a objetos. Por ejemplo, se pueden usar números reales para representar coordenadas espaciales, variables categóricas para clases sociales, números enteros para edades, etc.

Para modelar la dispersión de una enfermedad es apropiado construir ABMs epidemiológicos tomando una población de agentes en el que cada uno de ellos representa a una persona o individuo susceptible a contraer y/o transmitir la enfermedad (Roche et al., 2011). Incluso, un agente puede representar a un grupo de individuos de características similares. Utilizando variables categóricas se puede etiquetar a cada agente con su status o categoría epidemiológica (susceptible, infectado, etc.) y estas etiquetas pueden ser cambiadas cuando existe un contacto infeccioso o a medida que la enfermedad se desarrolla en el tiempo. El modelado de los contactos

entre agentes admite múltiples representaciones, se pueden hacer mezclas al azar, a través de grupos, con redes de contactos o modelando explícitamente la movilidad de los agentes en el espacio. El uso de ABMs para epidemiología tiene la virtud de que permite modelar de una manera explícita, y con gran detalle, características relevantes del sistema que suceden en la micro-escala de la interacción entre individuos. Por ejemplo, la disminución de la inmunidad puede ser representada en cada individuo a través de un atributo. Es posible modelar de manera natural medidas de control como cuarentenas, rastreo de contactos o efectos de vacunación (Silva et al., 2020). Muchos modelos de ecuaciones diferenciales asumen que dentro de cada compartimento la mezcla entre individuos es homogénea; para ABMs epidemiológicos, al disponer de caracterizaciones de cada individuo, es posible obtener patrones de interacción mas ricos a través de redes de contacto o el mecanismo que resulte más conveniente. De esta manera se pueden capturar efectos difíciles de representar con modelos de ecuaciones diferenciales. A pesar de disponer de información de cada agente, en la simulación con ABMs el interés está en el comportamiento global del sistema. Se suele contar con una función que de alguna manera resuma al estado de la población como un todo a través de un conjunto de variables que llamaremos variables agregadas. En general, si tenemos que la población de agentes es A, llamaremos $\mathbf{x}$ a las variables agregadas y ϕ a la función sintetizante o agregante que realiza el mapeo, de manera que:

$$\phi(A) = \mathbf{x} \tag{5.1}$$

El comportamiento de las variables agregadas emergerá del comportamiento a nivel individual del ABM. En el caso de modelos epidemiológicos es razonable, por ejemplo, utilizar como resumen el conteo de la cantidad de individuos en cada categoría para saber cuantos susceptibles, infectados, recuperados, etc. hay.

Con el surgimiento de la pandemia de COVID-19 se han desarrollado una gran diversidad de ABMs. Algunos de estos incorporan, entre otras características, estructura de edades y redes de contactos a través de escuelas, casas, lugares de trabajo que permiten modelar de manera más realista las mezclas que dan lugar a los contactos (Flaxman et al., 2020; Kerr et al., 2020; Simoy y Aparicio, 2021). Además de que el aumento de la capacidad de cómputo hace más viable la utilización de ABMs, el gran caudal de datos recolectados a nivel individual a través de dispositivos electrónicos es otro gran aliciente para el incremento en su popularidad. En Aleta et al., 2020 se utilizan datos demográficos y de movilidad para construir las redes de contactos y distribución de hogares en un ABM que permite evaluar los efectos de las intervenciones no-farmacéuticas.

A pesar de la flexibilidad y expresividad que permite el modelado mediante agentes, sigue siendo necesario calibrarlos adecuadamente. En general, el aumento en complejidad viene acompañado de un aumento en el número de parámetros, los cuales no siempre son especificables a través del conocimiento que se tenga sobre el sistema. Ha habido avances en el desarrollo de técnicas de inferencia para estimar parámetros en ABMs utilizando observaciones. Estos están principalmente orientados a la maximización de la verosimilitud. En particular en Hooten et al., 2020 se discuten una variedad de métodos para maximizar aproximaciones de la verosimilitud usando cómputo Bayesiano aproximado (o ABM por *Approximate Bayesian Computation*) junto con técnicas de MCMC (*Markov Chain Monte Carlo*).

Como los ABMs se enmarcan en un contexto de sistemas con evolución temporal parcialmente observados, es posible pensar que la caja de herramientas provista por la asimilación de datos secuencial es potencialmente de utilidad para realizar

las tareas de inferencia. Un trabajo pionero en esta dirección es Ward et al., 2016 en el que se utiliza el EnKF para asimilar datos de cámaras en calles de Leeds con un modelo de agentes para estudiar el comportamiento del tráfico peatonal. A pesar de obtener resultados satisfactorios, se señala la dificultad asociada a la sensibilidad respecto a parámetros del modelo y la necesidad optimizar el código para modelos de gran tamaño. Como mencionamos anteriormente, el interés no necesariamente está puesto en el estado de cada individuo sino en el estado global del sistema o de un conjunto de variables que den una descripción de la población de agentes que se considere relevante. Además, a pesar de que el estado a la escala microscópica determina de manera total al sistema y que el objetivo de la inferencia fuera, idealmente, representar esta escala de la manera más precisa posible, sucede que usualmente las observaciones que se tienen del sistema son de las variables sintetizantes, es decir de la escala macroscópica. Esto significa que las observaciones pueden no tener suficiente "resolución" como para que se pueda determinar el estado de los atributos de cada agente. Nuestra propuesta, publicada en Cocucci, Pulido, Aparicio et al., 2022, es la de asimilar datos sobre las variables agregadas mediante técnicas basadas en ensambles. El mapeo de la escala microscópica a macroscópica está dado por la función sintetizante ϕ pero no tenemos una representación para su inversa. De hecho ϕ puede no ser inyectiva y por lo tanto no-invertible: muchas configuraciones de los agentes pueden resultar en el mismo estado de las variables agregadas. Esta situación debe considerarse, a la hora de asimilar observaciones, para hacer un emparejamiento entre los estados macroscópicos observados y el estado de los atributos de los individuos. En este capítulo presentamos un esquema general de aplicación de métodos de asimilación de datos por ensambles para ABMs y dos implementaciones particulares para un ABM epidemiológico que especificamos a continuación.

5.2 Modelo epiABM

Aquí especificamos un ABM epidemiológico diseñado para modelar la dinámica de contagio de enfermedades infecciosas, en particular para el COVID-19 y que denominaremos epiABM. Este es un modelo sencillo para evaluar el rendimiento de la metodología para aplicar asimilación de datos por ensambles en ABMs. Si se quisieran realizar predicciones o asesorar en toma de decisiones, este debería ser adaptado para ese propósito incorporando otras características (por ejemplo, el surgimiento de nuevas cepas y mayor complejidad en la movilidad de agentes).

En el modelo epiABM, el estado de salud de cada agente está descripto por una etiqueta para representar una entre siete categorías. Tenemos a los susceptibles a contraer la enfermedad (S). También consideramos a los expuestos a la enfermedad pero que aún no son infecciosos debido a que el virus está incubando (E). Después de estar expuestos pasan a una de dos posibles clases de infecciosos. Los infectados leves (I_M) son los que desarrollan una sintomatología que no requiere hospitalización y que eventualmente se recuperan. Los infectados graves (I_S) son los que requerirán hospitalización. Por su parte, los hospitalizados pueden recuperarse o morir. La categoría R denomina a los recuperados y la D a los muertos. Asumimos que los recuperados adquieren inmunidad permanente, lo cual no es una suposición realista para simulaciones a largo plazo de COVID-19 puesto que las reinfecciones son posibles. El diagrama de la Figura 5.1 representa de manera esquemática la progresión de la enfermedad a través de estos estadios.

En cada paso temporal, que consideraremos de un día, cada agente tendrá un número de contactos muestreado de una distribución Poisson de parámetro λ. Los

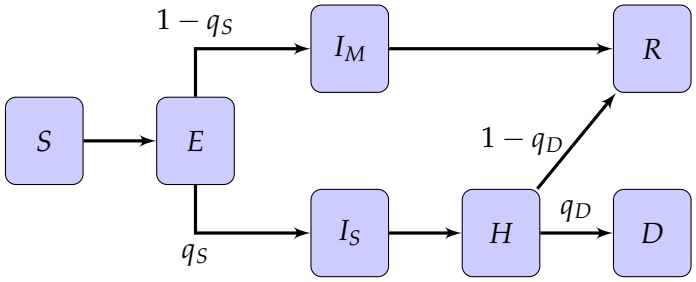

Figura 5.1: Diagrama de las categorías epidemiológicas del modelo epiABM.

agentes susceptibles podrán resultar infectados por un contacto con un agente infeccioso. El tiempo de permanencia de un agente en las categorías "intermedias" (E, I_M, I_S, H) está muestreado de una distribución Gamma. Si nombramos a este tiempo τ_c con $c \in \{E, I_S, I_M, H\}$ entonces $\tau_c \sim \Gamma(k_c, \theta_c)$ donde k_c y θ_c son los parámetros de forma y escala respectivamente. Con esta parametrización, la media μ_c y la varianza σ_c^2 cumplen con las relaciones $\mu_c = k_c\theta_c$ y $\sigma_c^2 = k_c\theta_c^2$. El tiempo τ_c es muestreado para cada agente cuando este entra a la categoría c y cuando este tiempo expira, el agente pasa a la siguiente categoría. Cuando un agente sale de la categoría E, tiene una probabilidad q_S de tener una enfermedad severa y $q_M = 1 - q_S$ de que su enfermedad sea leve. Análogamente, los hospitalizados tienen una probabilidad q_D de morir y $q_R = 1 - q_D$ de recuperarse.

Además de la estructura dada por el estatus epidemiológico, incluimos información demográfica y geográfica. El modelo considera una ciudad dividida en N_{loc} comunas. Cada agente vive en una casa (solo o con otros agentes) en una determinada comuna. El tamaño de los hogares sigue una distribución determinada por el vector de probabilidades p_H en el cual la i-ésima entrada determina la probabilidad de que un hogar tenga i miembros. Los contactos entre agentes pueden entonces ser definidos en base a esta estructura sencilla. Diferenciaremos entre contactos domésticos y casuales. Cada uno de los contactos diarios de los agentes tiene una probabilidad q_C de ser casual y $1 - q_C$ de ser doméstico. Los contactos domésticos se dan solamente entre miembros de un mismo hogar mientras que los casuales pueden ser, potencialmente con cualquier otro agente. Llamamos β_D (respectivamente β_C) a la probabilidad de que un contacto doméstico (respectivamente casual) sea infeccioso. De esta manera podemos escribir una expresión para el valor esperado de la probabilidad de infección global como $\beta = q_C\beta_C + (1 - q_C)\beta_D$. Por su parte, los contactos casuales van a estar mediados por la conectividad entre los diferentes barrios. Llamaremos C_{ij} a la probabilidad de que un contacto casual de un agente del barrio j se de con un agente del barrio i. Esto resulta en una matriz C de tamaño $N_{loc} \times N_{loc}$ que nombraremos matriz de contacto. Los elementos de la diagonal corresponden a la probabilidad de que un contacto casual se de entre habitantes del mismo barrio mientras que los elementos fuera de ella tienen que ver con los contactos entre habitantes de barrios distintos. La matriz C codifica entonces la movilidad entre barrios que a su vez se relaciona a las actividades sociales y laborales de la ciudad. El diseño de esta matriz puede incluir diferentes características de la estructura social y geográfica de la ciudad. Por ejemplo, sería esperable que los agentes transiten con más frecuencia su propio barrio, lo que implicaría valores más altos en los elementos de la diagonal y más pequeños por fuera de esta. Por otro lado, en caso de tener un barrio con más tránsito, por ejemplo un barrio céntrico, los valores correspondientes por fuera de la diagonal serían mayores que en el caso de un barrio

menos visitado. Para este trabajo consideraremos que C está fija en el tiempo pero existe la posibilidad de diseñar esta matriz con mayor detalle: por ejemplo se podría ajustar utilizando datos de movilidad que varíen en el tiempo para contemplar cambios que afecten al contacto entre personas o incluso para incluir los efectos de medidas de control como cuarentenas o restricciones de movilidad. Aquí utilizamos expresiones sencillas para la matriz de contactos porque el objetivo principal es el de la evaluación de la metodología propuesta para asimilar datos en ABMs. Una parametrización por defecto para el modelo se especifica en el Apéndice C.1.

El modelo puede ser extendido añadiendo más atributos a los agentes o representando con mayor detalle los patrones de contacto entre personas. Por ejemplo, sería posible incorporar edades o perfiles sociales para enriquecer la configuración de relaciones entre contactos que sean relevantes para la descripción de la dispersión de la enfermedad. Añadir eventos sociales de gran concurrencia, como por ejemplo lugares de trabajo o escuelas puede ser útil para el modelado de los fenómenos de dispersión masiva (conocidos como eventos *superspreader*). También es posible incluir otras características como las campañas de vacunación o el surgimiento de nuevas cepas del virus. Incorporando campos a la descripción del estado de los agentes se puede determinar si están vacunados, cuantas dosis recibieron, etc. Por otro lado, se podría indicar con qué variante del virus se contagiaron los agentes infectados.

El ABM subdivide al total de agentes en subpoblaciones de acuerdo a categorías epidemiológicas de manera similar a modelos compartimentales de ecuaciones diferenciales. Sin embargo los ABMs no pueden ser descriptos de manera directa con ecuaciones diferenciales. Cuando el número de agentes es grande las variables agregadas pueden suavizar el efecto de las componentes estocásticas (por ejemplo de los tiempos muestreados de distribuciones Gamma) y como resultado pueden llegar a ser reproducibles con modelos de ecuaciones diferenciales. Los ABMs tienen características, tal como el comportamiento resultante de tener agentes que residen en hogares y con tasas de infección diferenciadas entre contactos domésticos y casuales, para las cuales no está claro cómo se podrían traducir a un modelo basado en ecuaciones. El modelado a nivel microscópico de los ABMs puede tener consecuencias en la gran escala que pueden no ser fáciles de predecir. A pesar de esto, es importante destacar que los ABMs pueden tener un alto costo computacional lo cual ha motivado el uso de modelos de ecuaciones que actúan como sustitutos de un ABM y que pueden ser utilizados para realizar inferencia con bajo costo computacional (Hooten et al., 2020).

5.3 Asimilación de datos en ABMs

Para aplicar técnicas de asimilación de datos a ABMs es necesario entender primero que el estado propiamente dicho de un sistema de agentes en un instante de tiempo t está dado por el valor de los atributos de cada uno de los agentes que componen la población. En principio no es conveniente aplicar asimilación de datos en un espacio de estas características porque, por un lado, la cantidad de variables es enorme (del orden de la cantidad de agentes multiplicado por la cantidad de atributos) y por otro lado muchos de los atributos pueden no ser valores numéricos. Es posible aplicar filtros de partículas sobre espacios con variables categóricas pero apuntamos a usar el EnKF el cual trabaja sobre espacios de números reales. Incluso de ser factible asimilar datos en el espacio dado por los atributos de los agentes es probable que el interés principal esté en la inferencia sobre las variables agregadas. En el caso del modelo

epiABM las variables agregadas estarán dadas por la cantidad de agentes con determinados atributos (por ejemplo estado epidemiológico y comuna). Estos valores son enteros pero al ser lo suficientemente grandes pueden ser tratados por el EnKF como si pertenecieran a un espacio continuo. Además de este punto notemos que, en general las observaciones sobre el sistema serán sobre las variables agregadas, por lo que serán más informativas sobre el valor de estas cantidades y por lo tanto es más directo intentar la inferencia sobre el espacio que estas determinan.

Llamaremos al conjunto de valores de atributos del k-ésimo agente en una población a tiempo t como $A_{k,t}$. A su vez, al conjunto de estos valores para toda la población de agentes la llamaremos $A_t = \{A_{k,t}\}_{k=1}^{N_a}$, donde N_a es el número total de individuos. Debido a que nuestro enfoque apunta a utilizar técnicas basadas en muestras debemos considerar que tenemos no una población sino un ensamble de tamaño N_p de poblaciones de tamaño N_a. Este ensamble puede ser representado como $\{A_t^{(j)}\}_{j=1}^{N_p}$. Por su parte, el proceso de asimilación de datos se dará sobre las variables agregadas, o macroscópicas, $\mathbf{x}_t$. Estas se pueden obtener de una población de agentes A_t mediante la aplicación de la función sintentizante, de manera que tenemos $\mathbf{x}_t = \phi(A_t)$. De esta manera tenemos el ABM para evolucionar temporalmente a la población de agentes de A_t a A_{t+1} y mediante ϕ podemos obtener las variables agregadas $\mathbf{x}_{t+1}$. El inconveniente es que la función ϕ no será en general inyectiva y por lo tanto no será invertible, es decir que muchas configuraciones distintas de la población de agentes pueden dar las mismas variables agregadas. La técnica de asimilación de datos actualiza las variables sobre las que actúa cuando incorpora la información de una observación. En nuestro caso se actualizará a $\mathbf{x}$ pero como ϕ no es invertible no tenemos una manera explícita de dar cuenta de este cambio sobre la población de agentes. Es importante notar que tanto para el EnKF como para el filtro de partículas el modelo es tratado como una caja negra por lo que estas técnicas actuarán ignorando por completo la existencia del mapeo ϕ.

Al tiempo t supongamos que tenemos un ensamble $\{A_t^{f(j)}\}_{j=1}^{N_p}$. Cada miembro del ensamble $A_t^{f(j)}$ representa una población de agentes donde el supraíndice f indica que la muestra es de un *forecast*. Para obtener un pronóstico de las variables de estado simplemente debemos aplicar ϕ a cada $A_t^{f(j)}$ y de esa manera conseguimos el ensamble $\{\mathbf{x}_t^{f(j)}\}_{j=1}^{N_p}$. Utilizando este pronóstico en combinación con la observación $\mathbf{y}_t$ obtenemos, mediante la actualización de la técnica de asimilación de datos, un ensamble de análisis $\{\mathbf{x}_t^{a(j)}\}_{j=1}^{N_p}$. En este punto, en un esquema de asimilación de datos por ensambles tradicional utilizaríamos el modelo para evolucionar $\{\mathbf{x}_t^{a(j)}\}_{j=1}^{N_p}$ en $\{\mathbf{x}_{t+1}^{f(j)}\}_{j=1}^{N_p}$, es decir que con las variables de estado de análisis a tiempo t obtendríamos un pronóstico a tiempo $t+1$. Pero el ABM no puede transformar directamente $\mathbf{x}_t$ en $\mathbf{x}_{t+1}$ sino que evoluciona a A_t en A_{t+1}. Necesitamos entonces una representación de agentes $A_t^{a(j)}$ de las variables $\mathbf{x}_t^{a(j)}$. Esta población de agentes puede entonces ser usada por el ABM para obtener las poblaciones de agentes de pronóstico $\{A_{t+1}^{f(j)}\}_{j=1}^{N_p}$ para el tiempo siguiente y retomar la iteración pronóstico-análisis. En la Figura 5.2 se sintetiza este procedimiento general. El ingrediente faltante entonces es el ajuste de la población de agentes a las variables agregadas actualizadas mediante la información observacional. Para esto proponemos dos metodologías distintas para el caso del epiABM.

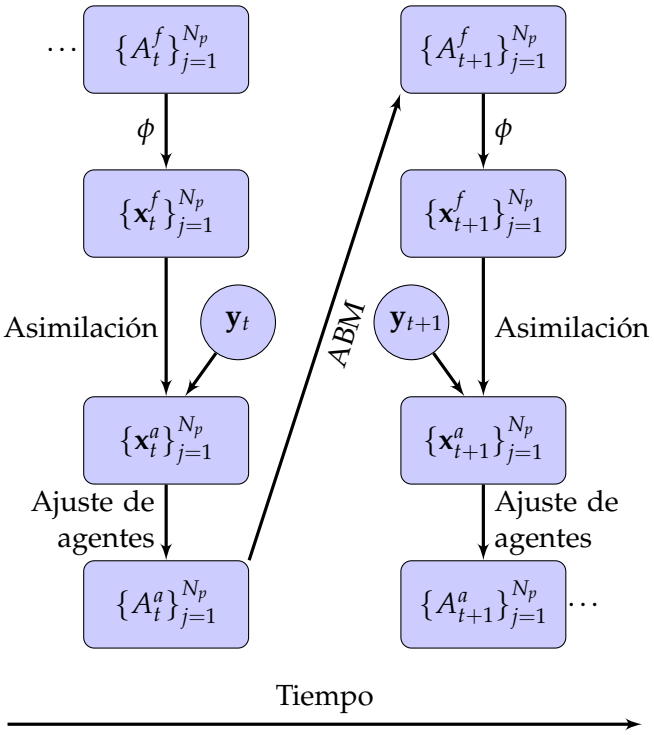

Figura 5.2: Esquema general de asimilación de datos por ensambles para ABMs

5.3.1 Metodologías de ajuste de agentes

Para aplicar el esquema propuesto es necesario entonces construir una población de agentes consistente con el estado de las variables macro una vez que estas fueron actualizadas por el sistema de asimilación de datos. El problema es que para cada $j = 1, ..., N_p$ tendremos una discordancia entre las variables de estado de la población de agentes que tenemos como pronóstico $\mathbf{x}_t^{f(j)} = \phi(A_t^{f(j)})$ y el valor corregido para las variables agregadas actualizadas $\mathbf{x}_t^{a(j)}$. Nuestra aproximación al problema es utilizar la misma población que se tenía como pronóstico, es decir $\{A_t^{f(j)}\}_{j=1}^{N_p}$, y hacer los cambios necesarios para obtener $\{A_t^{a(j)}\}_{j=1}^{N_p}$ de manera consistente con $\{\mathbf{x}_t^{a(j)}\}_{j=1}^{N_p}$. Esto está inspirado en que los agentes del análisis serán una corrección de los del pronóstico. Mientras mejor sea el pronóstico, menos agentes deberán ser modificados. Sin embargo, la estructura interna de los agentes puede ser muy compleja, con muchos atributos más allá de los considerados por la función sintetizante ϕ. Si estos atributos podrán ser ajustados de manera realista dependerá del ABM y de cuánto de la estructura interna de los agentes se ve representada en las variables del macro-estado.

Para el modelo epiABM presentado en la Sección 5.2 consideraremos dos metodologías distintas para corregir las poblaciones de agentes. Nuestra función sintetizante ϕ simplemente contará la cantidad de agentes en cada categoría epidemiológica en cada barrio por separado. Como tenemos $N_x = 7$ categorías epidemiológicas distintas y N_{loc} barrios, la dimensión de estas variables será $N_x N_{loc}$. Para la primera metodología, consideraremos las categorías donde "sobran" agentes (comparando

con el valor deseado, es decir el análisis de las variables agregadas) y los redistribuiremos en las categorías que muestran un déficit de agentes hasta lograr el objetivo de coincidir con los valores de $\mathbf{x}_t^{a(j)}$. Los agentes se seleccionan al azar (de las categorías correspondientes) y por ello llamamos a este método "redistribución aleatorizada". El cambio de categoría de agentes se hace mediante un cambio de etiquetas y el procedimiento se repite de manera independiente en cada barrio. Pero no sólo es necesario realizar cambios en las categorías epidemiológicas sino que hay otros atributos que deben ser atendidos. Por ejemplo, cada agente en las categorías E, I_M, I_S o H posee un contador que indica cuanto tiempo le queda para pasar a la categoría siguiente. Estos contadores son originalmente muestreados de distribuciones Gamma, como se especificó en la Sección 5.2. Entonces, cuando un agente es introducido por la actualización a una de estas categorías se debe dar un valor a este contador. Nuestra elección es muestrear el valor al azar de los agentes que ya están en dicha categoría para intentar preservar la distribución de ese valor dentro de la población. Aunque esta metodología es particular para la aplicación sobre nuestro modelo se pueden hacer implementaciones similares de manera general siempre y cuando tengamos un conocimiento *a priori* sobre la distribución del valor del atributo a través de la población de agentes. La idea del método es ser lo menos intrusivo posible con la población de agentes original y en este sentido la cantidad de individuos que se modifica es la mínima posible para lograr la coherencia con las variables agregadas.

La segunda metodología de ajuste de agentes no apunta a minimizar la cantidad de individuos modificados sino a intentar preservar la historia de cada uno de ellos. Para seleccionar los agentes a ser modificados intentaremos escoger aquellos que más probablemente tengan que cambiar de categoría según un criterio epidemiológico. Los cambios sólo se darán entonces entre categorías adyacentes en la cadena de progresión de la enfermedad representada en el diagrama de la Figura 5.1. Las correcciones comenzarán en las últimas etapas (R y D) y se avanzará hacia atrás hasta llegar a S. Además, la selección de los agentes que cambiarán de categoría no será al azar. Si un cambio debe ser hecho en la dirección del flujo del diagrama se elegirán los agentes que hayan pasado más tiempo en su categoría actual como los que más probablemente deban proseguir a la siguiente para hacer la corrección. De manera análoga, si un cambio debe ser hecho en el sentido contrario al flujo del diagrama se elegirán los agentes que hayan pasado menos tiempo en su actual categoría y se los retornará a la categoría anterior. La idea es preservar en lo posible las trayectorias individuales de cada agente. Para los cambios entre susceptibles y expuestos usaremos un criterio distinto, porque que un agente haya estado mucho tiempo como susceptible no significa que necesariamente sea más propenso a infectarse. En ese caso el criterio es mover de susceptibles a expuestos los agentes que más contactos riesgosos hayan tenido (es decir contactos con agentes infecciosos, los cuales no resultan siempre en infección). En la dirección opuesta, es decir para mover agentes de la categoría de expuestos a susceptibles se sigue usando el criterio de días de permanencia en la clase. Cada vez que un agente cambia de categoría el contador correspondiente se reinicia muestreando valores de los contadores de los agentes que ya estaban en la categoría donde fue reasignado. Este proceso se repite para cada barrio de manera independiente al igual que en la redistribución aleatorizada. Llamamos "redistribución con cascada hacia atrás" a la metodología resultante. Una desventaja del método es que cuando seleccionamos a los agentes con más permanencia en una categoría estamos truncando las colas de la distribución de ese atributo.

5.4 Evaluación experimental

Realizamos una evaluación experimental de los métodos propuestos acoplados con el EnKF en el modelo epiABM. Primero usamos observaciones sintéticas de manera que el estado verdadero del que son muestreadas las observaciones es generado con el modelo y está disponible para ser comparado con las estimaciones. Luego utilizamos datos de casos de COVID-19 en la Ciudad Autónoma de Buenos Aires (CABA), Argentina.

Como mencionamos anteriormente, la cantidad de variables de estado está dada por $N_x N_{loc}$ donde N_x es la cantidad de categorías epidemiológicas y N_{loc} la cantidad de barrios o comunas. En realidad, a estas variables de estado también se les agregan posiblemente los parámetros que se estimen mediante estado aumentado. A no ser que se mencione explícitamente utilizamos la parametrización por defecto especificada en el Apéndice C.1.

En los experimentos consideramos como variables observadas a los casos acumulados y muertes acumuladas, ambas desagregadas por barrio. Los casos acumulados pueden ser computados como la suma de I_M, I_S, H, R, D en cada barrio. Al error en las observaciones lo suponemos Gaussiano e insesgado con varianza proporcional a la variable observada correspondiente.

La versión del EnKF que utilizamos es la estocástica o con observaciones perturbadas (Burgers et al., 1998). Aunque es usual utilizar algún tipo de inflación multiplicativa o aditiva en este tipo de métodos para evitar el colapso del ensamble (ver Sección 2.4.2), pudimos detectar en experimentos preliminares que esto no se hace necesario debido a que la estocasticidad intrínseca del modelo mantiene a las trayectorias del ensamble con suficiente dispersión. Es decir que la aleatoriedad en las mecánicas en el microestado actúa como una fuente de error de modelo estocástico para las variables agregadas. Por otra parte, la variabilidad inicial del ensamble está dada por muestreo aleatorio en la inicialización de los atributos de los agentes.

Realizamos un análisis de sensibilidad en un experimento sintético para determinar un tamaño de ensamble adecuado y obtuvimos que con ensambles mayores a 50 miembros no se obtenía una mejora significativa en el error cuadrático medio (RMSE). Por otro lado, la varianza se estabiliza recién con ensambles de 100 miembros. Tomamos entonces 100 miembros de ensamble como el tamaño para los experimentos sintéticos. Para los experimentos con datos reales no podemos hacer la misma evaluación de la sensibilidad porque no contamos con las trayectorias reales, pero como tenemos aproximadamente el cuádruple de variables de estado (porque tenemos más cantidad de barrios) consideramos 400 miembros de ensamble. En ambos casos, incrementar el tamaño del ensamble no resulta en cambios detectables en los resultados.

5.4.1 Observaciones sintéticas

Hacemos aquí una evaluación experimental del métodos para distintas configuraciones simulando las observaciones. En la mayor parte de los experimentos, el modelo que se emplea para generar las observaciones es el mismo que se utiliza para realizar la inferencia, lo que permite comparar las trayectorias reales con las estimadas. Sin embargo en algunos casos consideramos parámetros desconocidos o incluso especificaciones distintas del modelo. Tomamos un total de $N_{loc} = 4$ barrios

y usamos la siguiente matriz de contacto:

$$C = \begin{pmatrix} 0.43 & 0.14 & 0.14 & 0.29 \\ 0.14 & 0.43 & 0.14 & 0.29 \\ 0.14 & 0.14 & 0.43 & 0.29 \\ 0.14 & 0.14 & 0.14 & 0.57 \end{pmatrix}.$$

Consideramos que los agentes tienen contactos casuales con mayor probabilidad dentro de sus propios barrios, lo que explica que los valores en la diagonal sean mayores. Adicionalmente, el barrio con mayor indexación es considerado como un barrio más concurrido por lo que tenemos mayores valores en la última columna.

Número de contactos variable

El número de contactos que tiene un agente por día está codificado a través del parámetro λ. En este experimento consideramos que el valor real de este parámetro disminuye linealmente en el tiempo y en el sistema de inferencia lo asumimos desconocido y lo estimamos mediante estado aumentado.

La Figura 5.3 muestra los resultados de las variables de estado reales y estimadas por el EnKF. Se puede ver que el ensamble estima correctamente a las trayectorias reales. El ensamble que estima las muertes acumuladas tiene muy poca varianza, esto se debe a que esta variable es directamente observada y con un error observacional relativamente pequeño. El resto de las variables no son observadas y sus estimaciones se logran a través de las correlaciones con las variables observadas de las que hace uso el EnKF. En la Figura 5.4 se grafica la evolución del valor real de λ junto con las trayectorias del ensamble. Las estimaciones capturan el cambio en el parámetro luego de un tiempo inicial de spin-up en que el sistema de asimilación sincroniza con las observaciones a medida que las incorpora. Alrededor del pico epidémico es cuando la varianza de las estimaciones es menor y luego de eso comienza a amplificarse. Esto se debe a que las observaciones son más informativas de λ (a través de correlaciones) cuando el número de infecciones es mayor. A medida que el número de infecciones disminuye la incerteza en las estimaciones comienza a crecer. La media del ensamble inicial se elije adrede en un valor distinto al valor real inicial de λ ya que el parámetro se supone desconocido. Por otro lado, consideramos una varianza inicial grande para que el ensamble pueda explorar mejor el espacio paramétrico. A medida que crece la correlación entre las variables observadas y el parámetro, esta varianza se comienza a encoger. En todos los casos se muestran los resultados para la redistribución aleatorizada pero estos son similares con el método de corrección de cascada hacia atrás.

La estructura dada por la distribución de agentes en hogares no es informada de manera directa por las observaciones puesto que estas sólo están diferenciadas por barrio. El sistema sin embargo mantiene una población de agentes por cada miembro de ensamble y es posible evaluar cómo se distribuyen las infecciones en los distintos tipos de hogares. Nuestra parametrización del modelo considera casas con un mínimo de un solo habitante hasta un máximo de cinco. En cada día de asimilación computamos qué proporción de infectados corresponde a cada tamaño de hogar: los resultados se muestran en la Figura 5.5. Cuando hay menos infecciones, es decir antes y después del pico epidémico las estimaciones tienen varianza más alta y son menos precisas mientras que durante el pico de infecciones la dispersión es mucho menor y las estimaciones más certeras. Podemos comparar la proporción de infectados en cada tipo de hogar con la proporción total de agentes residentes en

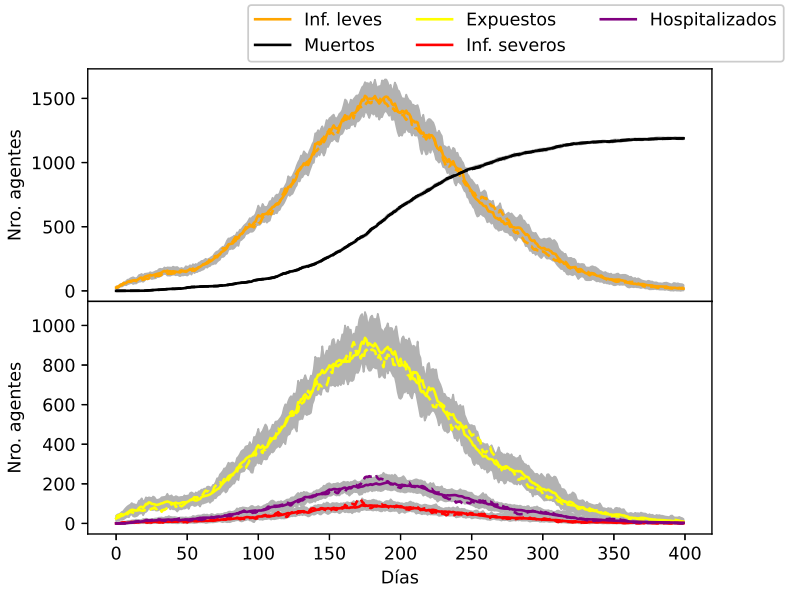

Figura 5.3: Estimaciones de las variables de estado I_M y D en el panel superior y de E, I_S y H en el inferior. En líneas continuas las medias, en gris los miembros del ensamble y en líneas intermitentes los valores reales.

cada tipo de hogar sin importar su estado epidemiológico. En casas más pequeñas la proporción estimada de infecciones es menor a la proporción de agentes que habitan en ese tipo de hogar. Lo opuesto sucede en casas de mayor tamaño. Esto se debe a que debido a los contactos domésticos la infección se dispersa con mayor rapidez en hogares con más habitantes. A su vez, la proporción de infectados en hogares más grandes es mayor al comienzo de la epidemia y menor hacia el final porque son los hogares que más rápidamente se saturan de inmunes (por infectarse más rápidamente).

Seguimiento de la microescala

Realizamos un experimento en el que evaluamos, en distintos niveles de agrupamiento, qué cantidad de agentes coinciden en su estado epidemiológico, comparando entre las estimaciones y la corrida real. El objetivo del experimento es comparar la evolución del estado en la micro-escala entre la población real de agentes y las de los miembros del ensamble. Contaremos la cantidad de agentes que tienen el mismo estado epidemiológico pero considerando diferentes agrupamientos de agentes. Cada agente tiene un número de identificación (id); la primera métrica consistirá en comparar el estado epidemiológico de los agentes id por id. La segunda métrica consistirá en comparar casa por casa para chequear coincidencias utilizando el id que posee cada casa. Para la tercera métrica agruparemos los agentes de acuerdo al tamaño de hogar en el que habitan. La última métrica agrupa agentes de acuerdo al tamaño de hogar y el barrio en que habitan. La única métrica que distingue agentes por su id es la primera, para el resto contamos las coincidencias simplemente contando la cantidad de agentes en la misma categoría epidemiológica. Estas métricas permiten evaluar el estado microscópico del sistema y sus desviaciones respecto al estado real. Utilizamos $N_{loc} = 4$ barrios y en lugar de utilizar un valor de λ global consideramos un valor distinto para cada locación, de manera que tenemos un vector de N_{loc} componentes $\lambda = (1.0 \;\; 0.8 \;\; 0.9 \;\; 0.7)$.

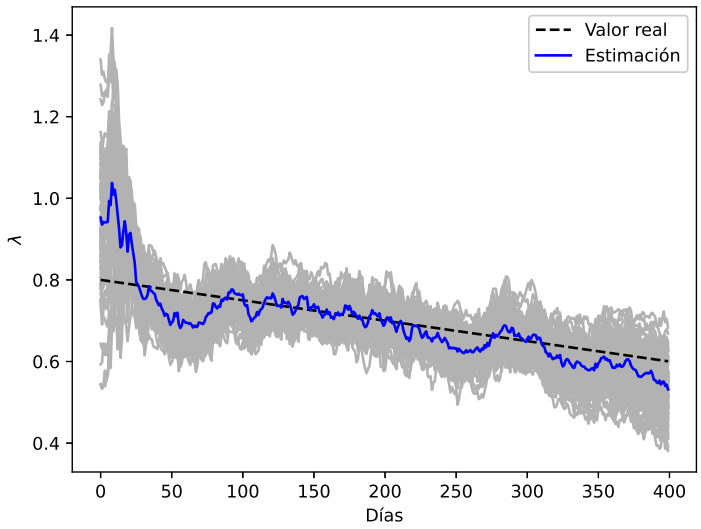

Figura 5.4: Estimaciones de λ. Media en azul y miembros del ensamble en gris. La línea intermitente indica el valor real.

La configuración inicial de los agentes para la corrida real y los miembros del ensamble es la misma por lo que al comienzo el porcentaje de coincidencias es 100% para todas las métricas. Para evaluar el impacto de la asimilación producimos además 100 simulaciones sin ningún tipo de asimilación para poder comparar los resultados. Estas serán diferentes a la trayectoria real debido a la estocasticidad del modelo.

La Figura 5.6 muestra los resultados para ambos métodos de ajuste de los agentes y para las simulaciones sin asimilación de datos. Cada panel muestra una de las métricas. La metodología de redistribución aleatorizada y de cascada hacia atrás producen resultados similares en todos los casos. Las trayectorias de las métricas para las corridas independientes tienen una gran dispersión mientras que cuando se asimilan datos estas están contenidas. Para las métricas de id de agentes y de casas, que son las que evalúan a una escala más microscópica, se puede ver que el porcentaje de coincidencias cae, se recupera un poco y finalmente se estabiliza tanto para las corridas con EnKF como para las simulaciones independientes. Estas últimas son ligeramente más consistentes con la corrida real en la primera etapa en que todas las métricas caen. Esto es posiblemente debido a que la metodología de ajuste debe modificar el estado de agentes para lograr la consistencia con las variables macroscópicas mientras que las simulaciones independientes mantienen en mayor medida la estructura inicial de la población de agentes.

En la escala de tamaños de hogares el EnKF mantiene un porcentaje alto de coincidencias (mayor al 90% para ambas metodologías de ajuste). En cambio, las simulaciones independientes muestran una gran variabilidad. El resultado para la métrica que agrupa por tipo de hogar y barrio da resultados similares. Algunas de las trayectorias de las simulaciones independientes tendrán picos epidémicos que no coinciden con el de la trayectoria real por lo que es esperable que las métricas para estas corridas decrezcan. El porcentaje de coincidencias luego se recupera puesto que el tamaño final de la epidemia será similar para la corrida real y las simulaciones y la mayor parte de los agentes estará en la categoría de susceptibles o recuperados. Por su parte, para el EnKF tenemos un buen nivel de coincidencias para estas métricas

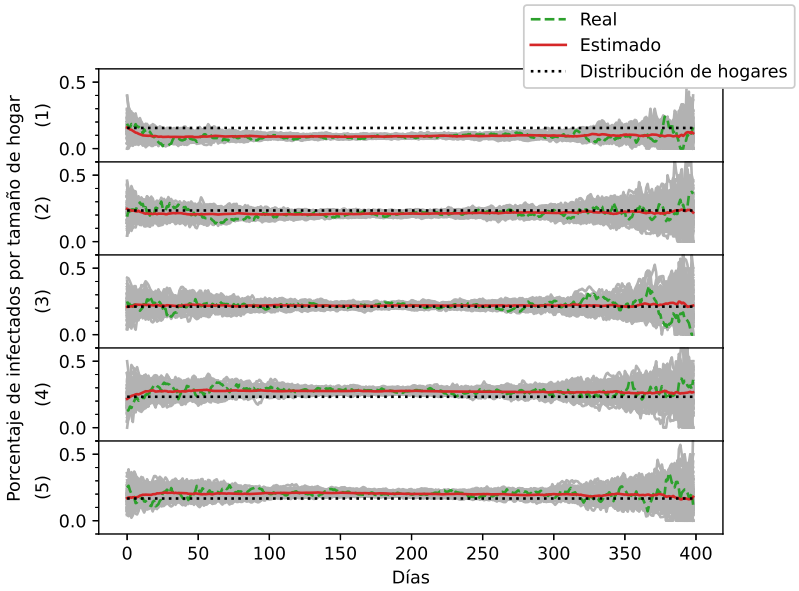

Figura 5.5: Proporción de infectados por tipo de hogar. Cada panel corresponde a distintos tamaños de hogar desde 1 habitante en panel superior a hogares de 5 habitantes en el panel inferior. Las líneas punteadas indican la proporción de habitantes en total (sin considerar estado epidemiológico) por cada tipo de hogar. Las líneas verdes señalan las proporciones de infectados de la corrida natural y las rojas las medias de las estimaciones mientras que las grises son las de los miembros de ensamble.

debido a que al asimilar datos, las trayectorias del ensamble pueden detectar el pico epidémico y sincronizar con el estado real en esta escala. Es importante notar que las observaciones en estos experimentos están separadas por barrio pero no son explícitamente informativas respecto a las infecciones en distintos tamaños de hogar. Aún así el porcentaje de coincidencias es bastante alto para estas métricas. La proporción de coincidencias considerando ids de agentes o agrupando por casas particulares es similar usando asimilación de datos o no haciéndolo. Esto es posiblemente porque los ids de agentes o de casas no juegan un rol particular en las dinámicas de contagio. En cambio, el tipo de hogar sí tiene un efecto en la dispersión de la enfermedad y por lo tanto este es capturado por el EnKF.

Estimación de casos no detectados

Es esperable que los datos de infecciones de COVID-19 reportados por los sistemas nacionales de salud estén afectados por el subreporte de casos ya que existen casos muy leves o incluso asintomáticos que no son contabilizados. Para evaluar si a través del sistema de asimilación podemos capturar los casos no reportados en las observaciones, modificamos levemente nuestro modelo y consideramos que las infecciones leves tienen una posibilidad q_A de ser asintomáticos y por lo tanto no estar contabilizados en los datos observacionales. Para el experimento utilizamos una configuración similar a la de los experimentos anteriores pero tomando un valor fijo de $\lambda = 0.8$ e introduciendo q_A en el estado aumentado. Para dar cuenta de los asintomáticos agregamos dos nuevas categorías epidemiológicas: los infectados asintomáticos I_A y los recuperados de una infección asintomática R_A. El diagrama de las categorías epidemiológicas se puede representar entonces como en la Figura 5.7.

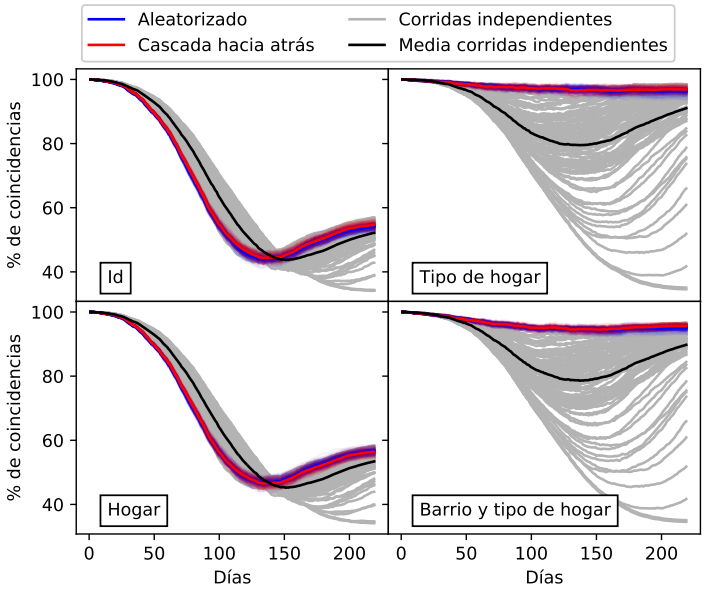

Figura 5.6: Proporción de aciertos (cantidad de agentes con coincidencia en el estado epidemiológico respecto a la corrida natural) para las 4 métricas distintas (una en cada panel). Las líneas grises corresponden a las corridas independientes y las líneas negras son sus respectivas medias. Las líneas rojas y azules con transparencias corresponden a los miembros de ensamble para las corridas con EnKF para el método de resdistribución aleatorizada (azul) y de cascada hacia atrás (rojo)

También sería posible simular un escenario similar considerando coeficientes desconocidos en el operador observacional $\mathbf{H}_t$ que serían estimados mediante estado aumentado.

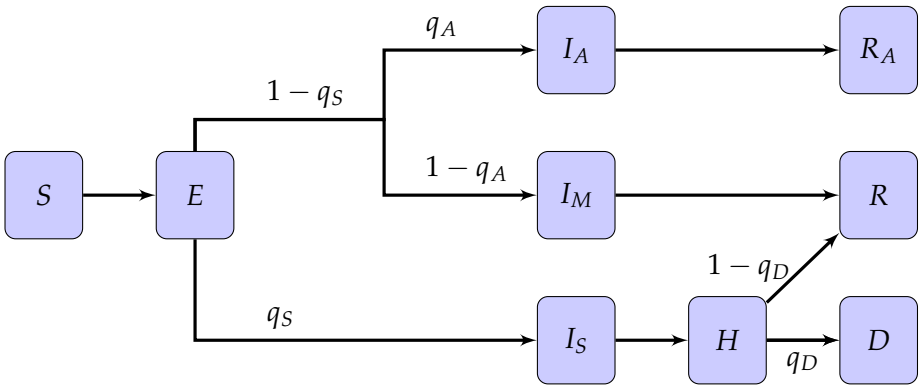

Figura 5.7: Diagrama para el epiABM con asintomáticos

Las variables observadas serán los confirmados acumulados como la suma de I_M, I_S, R y D y los muertos acumulados, D. Además incorporaremos una nueva variable observacional que llamaremos positividad global. Notemos que para los casos confirmados no incorporamos ni a I_A ni a R_A puesto que estos casos pasarían en principio inadvertidos para el sistema de reportes. La positividad global será el resultado de un testeo diario aleatorio a un porcentaje fijo de la población. Los tests darán positivos para I_M, I_S, I_A y H por lo que estas observaciones sí podrán capturar algo de información respecto a la cantidad de asintomáticos. Los tests son globales y

no están separados por locación por lo que dan una idea de la circulación global del virus. El error observacional de esta variable no será Gaussiano aditivo sino que será el error de muestreo que resulta de la aleatoriedad de los tests. Asumimos que un 1% de la población es testeada cada día. En la práctica los testeos no se suelen hacer completamente al azar sino que se hacen sobre casos posibles pero estos valores podrían ser utilizados para calcular aproximaciones de la positividad global. El valor de q_A será estimado mediante estado aumentado y se espera que las correlaciones de la positividad global con este parámetro permitan su inferencia.

La Figura 5.8 muestra los resultados para los infectados leves, los asintomáticos y la positividad global y el testeo. El comportamiento de los asintomáticos es capturado por el EnKF pero como es de esperarse las estimaciones de los infectados leves son más precisas. Esto es porque los asintomáticos son solo informados a través de la positividad global mientras que los infectados leves también son observados mediante los infectados acumulados por barrio. La discrepancia entre los asintomáticos y el valor real de esta variable está correlacionada con la discrepancia entre la positividad global y estimada. Esto sugiere que mientras más sepamos sobre la circulación general del virus, mejor podrá ser la inferencia sobre los casos no reportados. La Figura 5.9 muestra las estimaciones de la probabilidad q_A. Se ve en los primeros ciclos de spin-up una reducción de la varianza inicial y luego el ensamble sincroniza con el valor real. Cuando la positividad es menor, antes y después del pico epidémico, la varianza del ensamble es mayor que durante el pico epidémico, en donde las estimaciones son más precisas y de menos varianza. Este efecto se debe a que la positividad sólo será informativa de la proporción de asintomáticos en la medida que haya suficientes casos como para que las correlaciones entre los testeos y el parámetro sean lo bastante fuertes.

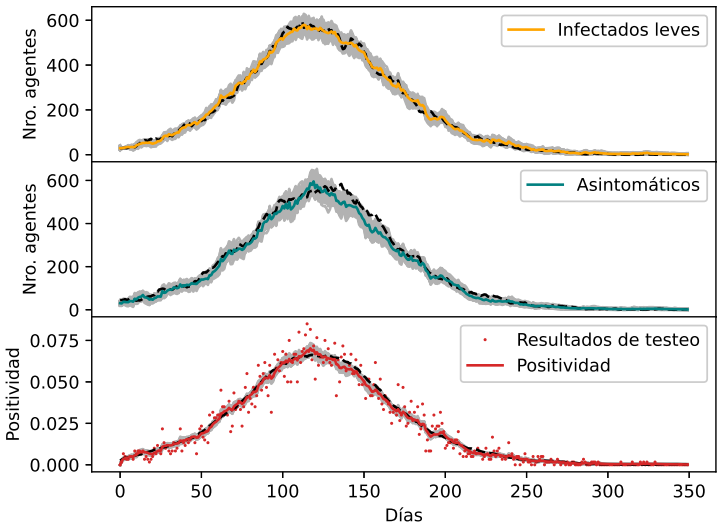

Figura 5.8: Estimaciones de infectados leves, asintomáticos y positividad global. Las líneas grises corresponden a los miembros del ensamble y las continuas a sus medias. Las líneas intermitentes indican los valores reales. En el panel inferior los puntos indican la positividad observada del testeo.

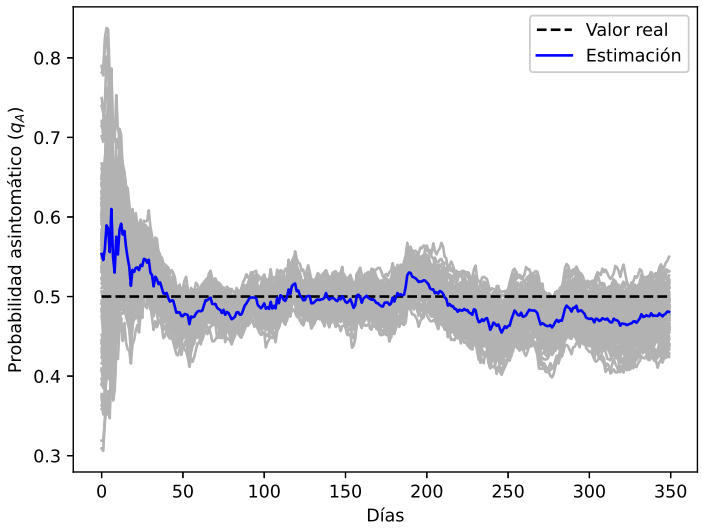

Figura 5.9: Estimaciones de la probabilidad q_A de que un infectado leve sea asintomático. Las líneas grises indican a los miembros del ensamble, y la azul a su media. El valor real se representa con la línea intermitente

Respuesta al error de modelo

En este experimento buscamos evaluar si la metodología es lo suficientemente robusta como para responder en situaciones en que el modelo está erróneamente especificado. Para ello diseñamos el experimento de manera que la corrida natural y las del EnKF tienen configuraciones diferentes. Para generar las observaciones sintéticas, el número de contactos diarios va a ser muestreado, no de una Poisson como hicimos hasta ahora, sino de una distribución geométrica con parámetro $p = 0.5$. Por el otro lado, para la inferencia con EnKF usaremos la distribución de Poisson como en los experimentos anteriores. Adicionalmente usaremos distintas distribuciones de hogares: para la corrida natural tomaremos $p_H = (0.33 \quad 0.27 \quad 0.2 \quad 0.13 \quad 0.07)$ lo que resulta en que las casas con más agentes son más infrecuentes que las que tienen menos habitantes, lo cual es esperable en un medio urbano. Para el EnKF utilizaremos esta distribución pero también repetiremos el experimento con una distribución uniforme de hogares $p_H = (0.2 \quad 0.2 \quad 0.2 \quad 0.2 \quad 0.2)$ y otra "desbalanceada" $p_H = (0.07 \quad 0.13 \quad 0.2 \quad 0.27 \quad 0.33)$ en la que los hogares de más habitantes son más frecuentes que los de menos. La utilización de distribuciones de hogares diferentes va a resultar en distintas dinámicas de propagación debido a la variación en la estructura de la red de contactos. El parámetro λ se estima mediante estado aumentado.

La Figura 5.10 muestra las trayectorias de las estimaciones de λ para los distintos escenarios de distribución de hogares. La distribución para el número de contactos para el EnKF es una Poisson mientras que para la corrida real es geométrica, por lo tanto la estimación de λ debería parametrizar al número de contactos diarios de manera que se parezca lo más posible a una geométrica. Para evaluar el parecido entre distribuciones entonces calculamos para una grilla de valores de λ la divergencia de Kullback-Leibler (KL) entre la Poisson y la verdadera distribución geométrica. Esta cantidad codifica cuánta información se pierde al utilizar la distribución Poisson respecto a la geométrica. En la Figura 5.10 también se muestra la divergencia KL junto con las estimaciones de λ. Estas están en regiones de baja divergencia KL: esto

significa que el sistema se auto-calibra buscando el valor de λ que mejor pueda reproducir las características del modelo original. La mejor estimación en términos de divergencia KL es para la corrida que utiliza la distribución real de hogares, lo cual es esperable. La distribución uniforme de hogares tiene más casas de mayor tamaño y la desbalanceada aún más. Esto causa una dispersión más rápida de la enfermedad debido a que habrá más contactos domésticos. El hecho de que las estimaciones de λ sean menores en estos casos se debe a que el sistema se auto-calibra y por lo tanto compensa por la distribución de hogares incorrecta que produce un mayor número de contactos domésticos (los cuales tienen mayor probabilidad de ser contagiosos).

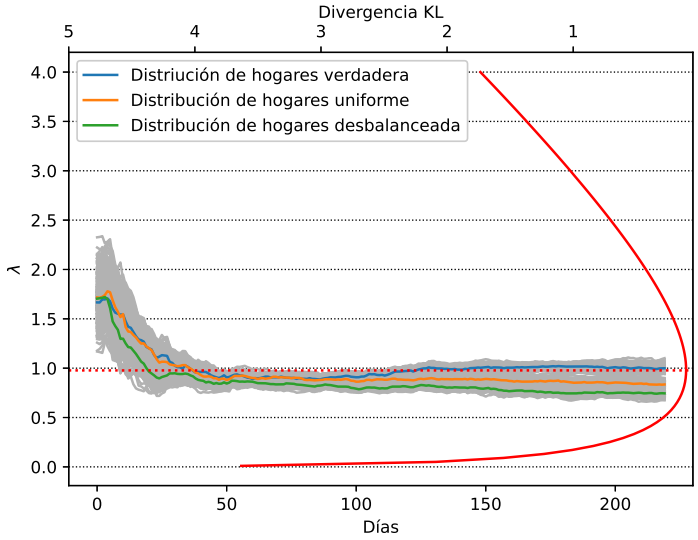

Figura 5.10: Estimaciones de λ para tres escenarios de distribución de hogares distintas. La curva roja representa a la divergencia KL y sus valores están en el eje superior del gráfico. Las líneas grises representan a miembros del ensamble y las líneas sólidas correspondientes a sus medias

5.4.2 Datos CABA, Argentina

Para evaluar el sistema con datos realistas realizamos un experimento con los datos de COVID-19 de CABA, Argentina, provistos por el Ministerio de Salud y publicados con actualizaciones diarias en https://data.buenosaires.gob.ar. La distribución de hogares y la población es tomada de datos censales disponibles en el mismo sitio. CABA se divide en 15 comunas. Tomaremos a cada una de ellas como un barrio para el modelo (a pesar de que cada comuna puede estar compuesta por más de un barrio). Los datos están agrupados de acuerdo a las distintas comunas. Utilizaremos como observaciones sobre el sistema a los infectados acumulados y las muertes acumuladas. Consideraremos que el parámetro λ es el mismo para todas las comunas y lo estimaremos mediante estado aumentado. Para la matriz de contacto consideraremos que la mitad de los contactos casuales son dentro de la misma comuna y la otra mitad están distribuidos de manera proporcional a la densidad poblacional de las otras comunas. Utilizamos $N_a = 3 \cdot 10^5$ agentes pero la población de CABA es aproximadamente $3 \cdot 10^6$ por lo que escalamos hacia abajo los datos en 10.

La Figura 5.11 muestra las estimaciones de λ. Como este parámetro codifica el número de contactos que un agente tiene por día es esperable que esté correlacionado con el número de infecciones confirmadas diarias y de hecho se puede ver

que tienen una tendencia similar. Esto sucede porque las probabilidades de infección, β_C y β_D, se consideran constantes durante la simulación por lo que los cambios en las infecciones diarias van a ser explicados por cambios en λ con el correspondiente desfasaje debido al tiempo de incubación. La evolución de la tasa de contactos estimada es consistente con los tres picos epidémicos en Argentina hasta mediados de 2021.

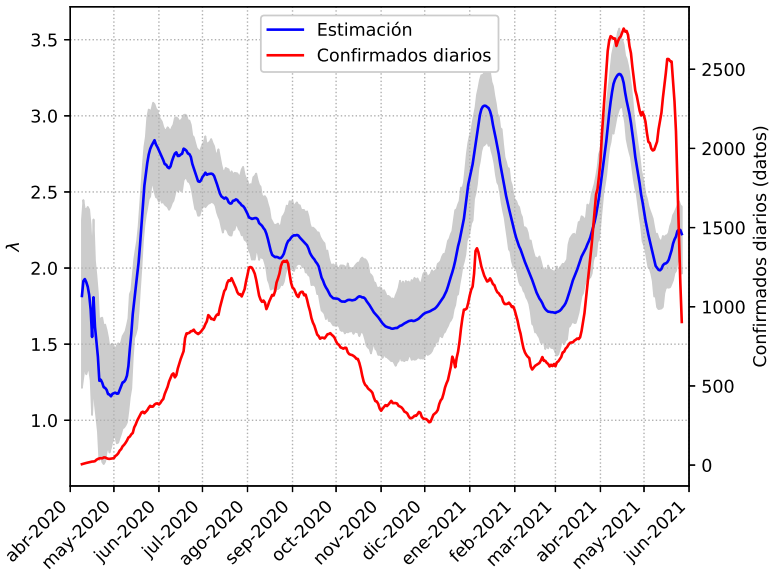

Figura 5.11: Estimaciones de λ junto a los casos diarios confirmados suavizados mediante una media móvil de 7 días (estos valores se representan en el eje derecho de la gráfica). Los miembros del ensamble se indican con líneas grises.

Es importante notar que sólo estimamos el parámetro λ dejando al resto fijo, cuando en realidad existen múltiples efectos que pueden influir en la propagación de la enfermedad. Esto significa que el ajuste del modelo a los datos se da a través de λ a pesar de que puede haber sido otro fenómeno el que provocó un determinado cambio en la tendencia de los datos. Por ejemplo, una disminución de los casos causada por el mayor uso de tapabocas debería ser representado mediante una baja en la probabilidad de infecciones casuales β_C pero, dado que este parámetro está fijo, el cambio será capturado por λ. A pesar de que esto puede ser impreciso, notamos que intentar estimar muchos parámetros con efectos similares en los datos pude llevar a una sobreparametrización y falta de identificabilidad.

La Figura 5.12 muestra algunas de las variables agregadas del sistema sumadas a través de todos los barrios. Las muertes acumuladas tienen menor varianza que el resto de las variables. De manera similar a los experimentos con observaciones sintéticas, esto se debe a que las muertes son directamente observadas mientras que las otras variables son observadas a través de la suma de varios estados epidemiológicos.

En la Figura 5.13 se puede ver el número de casos diarios. En los datos se puede ver el efecto del subreporte de los fines de semana pero las estimaciones suavizan en parte este efecto. La estimación entonces es más cercana a lo que esperamos que es el estado verdadero. La Figura 5.14 se muestran los mismos valores pero desagregados por comuna. Todas ellas tienen tendencias similares con algunas diferencias en la forma de los picos epidémicos pero estos efectos son capturados por el EnKF.

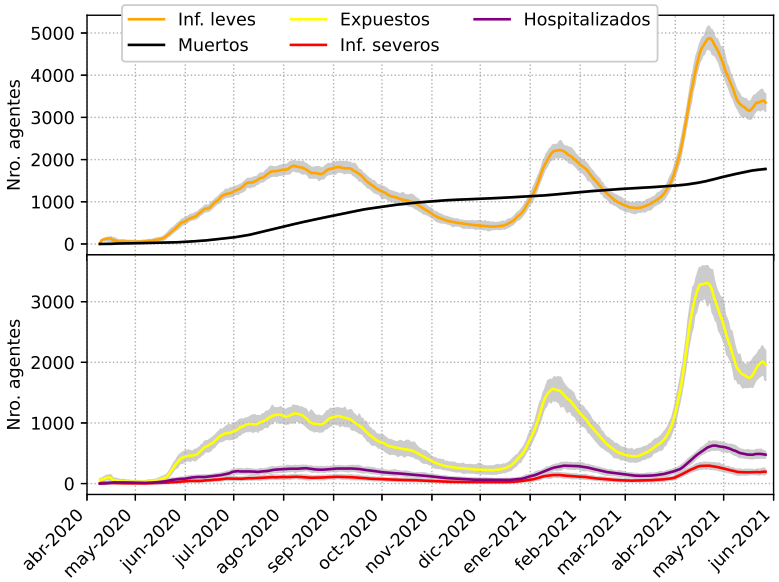

Figura 5.12: Estimaciones de las variables de estado I_M y D en el panel superior y de E, I_S y H en el inferior. En líneas continuas las medias, en gris los miembros del ensamble.

5.4.3 Discusión

En nuestro modelo epidemiológico, la evaluación experimental de las metodologías que propusimos para aplicar asimilación de datos por ensambles en ABMs da muestras prometedoras para la inferencia en este tipo de sistemas. Las metodologías de ajuste de agentes descriptas en la Sección 5.3.1 dan resultados similares y pueden recuperar los valores reales de las variables macroscópicas. En principio, las estimaciones sobre el estado microscópico de los agentes depende de la granularidad de las observaciones y de cuánto informen estas sobre la microescala. Por otro lado, los experimentos muestran que se puede sacar provecho del método de estado aumentado para la estimación de parámetros del modelo incluso en situaciones de especificaciones incorrectas de modelo como se muestra en los resultados de la Figura 5.10. En cuanto a su aplicación con datos reales, la correlación entre la incidencia diaria de casos y el parámetro λ que se puede ver en la Figura 5.11 sugiere que el sistema responde a los cambios del escenario epidemiológico que subyace en los datos. Es importante notar que en general, tanto en ABMs como en modelos de ecuaciones diferenciales, cuando se utilizan parámetros con efectos similares sobre las trayectorias observadas del sistema los efectos pueden no ser diferenciados por la técnica de asimilación. Esto incentiva la utilización de modelos parsimoniosos para poder sacar provecho de las herramientas de inferencia que aporta la asimilación de datos.

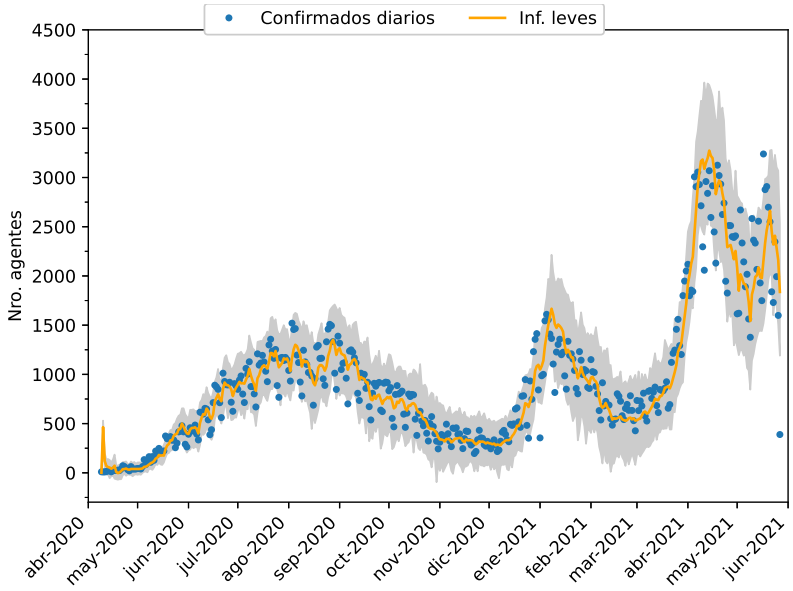

Figura 5.13: Estimaciones de infectados diarios en toda la ciudad. Las líneas grises indican miembros del ensamble y la línea sólida su media. Las observaciones se representan con puntos.

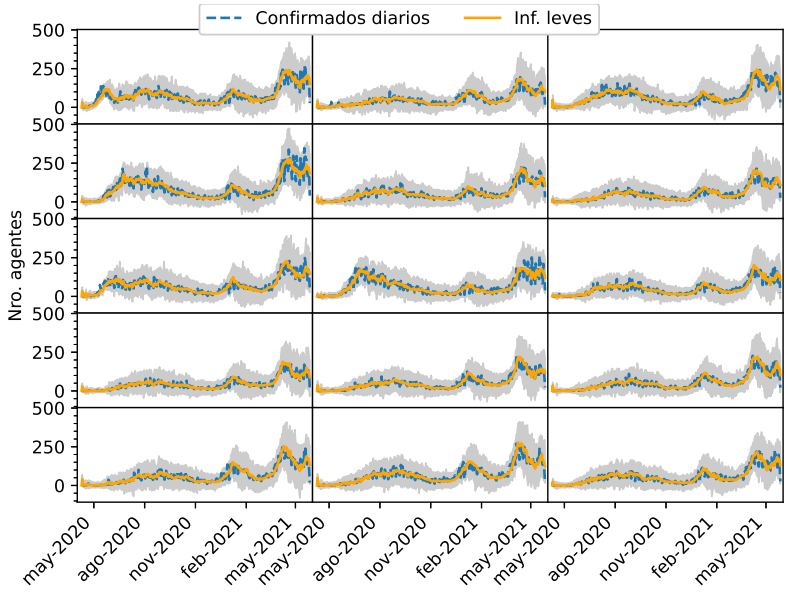

Figura 5.14: Estimaciones de infectados diarios en cada comuna de la ciudad. Las comunas están numeradas del 1 al 15 y están ordenadas en la gráfica de izquierda a derecha y de arriba a abajo. Las líneas grises indican miembros del ensamble y la línea sólida su media. Las observaciones se representan con líneas intermitentes.

Capítulo 6

Conclusiones

En este trabajo nos propusimos estudiar el tratamiento de errores en sistemas de asimilación de datos, sobretodo con técnicas basadas en ensambles, para luego explorar las posiblidades de aplicar metodologías de este tipo en sistemas epidemiológicos. Para esto dimos un marco teórico general de la asimilación de datos como problema de inferencia Bayesiana en modelos de Markov escondidos y proporcionamos un panorama de las técnicas más populares, dando especial atención a los filtros de partículas y los filtros de Kalman por ensambles. Luego presentamos el problema de la especificación de errores observacionales y de modelo y su impacto en la performance en sistemas de asimilación de datos. Mencionamos algunas de las estrategias que se han propuesto para la estimación de estas incertezas y nos focalizamos en el algoritmo EM. En este punto exponemos el desarrollo de lo que constituye una de las contribuciones principales del trabajo: una metodología *online* de estimación del error de modelo y observacional basada en el EM junto con una evaluación experimental de su desempeño en escenarios de interés para el área. Tiene la propiedad de acoplarse a técnicas de asimilación de datos por ensamble modernas como el EnKF o de filtros de partículas como el VMPF. También realizamos experimentos de aplicación del EM *online* para un modelo epidemiológico compartimental simple de COVID-19 utilizando tanto datos sintéticos como reales. Previo a esto presentamos los modelos epidemiológicos compartimentales clásicos así como los antecedentes de inferencia sobre éstos, haciendo énfasis en los trabajos que utilizan técnicas de asimilación de datos. El creciente interés en modelos basados en agentes nos motivó a extender la aplicación de asimilación de datos a este tipo de modelado. En este sentido, otra de las contribuciones más relevantes de este trabajo es la de un marco general de aplicación de técnicas de asimilación de datos por ensambles en modelos basados en agentes. Trabajamos sobre un modelo para COVID-19 y evaluamos experimentalmente la metodología en este contexto, sin embargo la estrategia tiene la generalidad suficiente como para ser aplicable a otros tipos de modelos basados en agentes.

El algoritmo EM *online* que desarrollamos mostró resultados satisfactorios en los escenarios en los que fue evaluado. La versión que se acopla al EnKF con un paso de suavizado de EnKS tiene una menor complejidad computacional que el EM *offline* acoplado al EnKS (Pulido, Tandeo et al., 2018). Además tiene la propiedad de que puede capturar cambios temporales lentos en los parámetros e incluso recuperar la estructura de covarianzas de las matrices que parametrizan los errores. La performance del método sin embargo depende de la elección de la tasa de aprendizaje. Esta cuantifica la importancia que se le da a la observación que está siendo procesada en comparación a la estimación previa. El problema es similar a los de la elección de una tasa de aprendizaje para métodos de descenso de gradiente en *machine learning*. Por lo tanto, una posible línea de investigación es buscar incluir en el EM *online* una tasa de aprendizaje adaptativa del tipo momentum, ADAGRAD,

ADADELTA o ADAM (Kingma y Ba, 2014; Zeiler, 2012). Estos métodos se adaptan de manera dinámica a la tarea de optimización y pueden mejorar dramáticamente la convergencia. Creemos que la maximización de la verosimilitud mediante EM puede ser reinterpretada como un método de gradientes lo que permitiría utilizar la vasta caja de herramientas provista en gran parte por el desarrollo moderno de métodos de *machine learning*. En particular, es muy posible que este enfoque permita una reinterpretación similar para los métodos *online*.

Otro desafío de interés para futuros trabajos es la extensión de la metodología a sistemas de alta dimensionalidad. El problema de la asimilación de datos en alta dimensionalidad atrae gran interés debido a que muchos modelos geofísicos consideran grillas espaciales que queden alcanzar las 10^9 dimensiones. Las representaciones de distribuciones mediante muestras en altas dimensiones se deteriora debido a la maldición de la dimensionalidad. El EnKF y el VMPF mitigan en parte el problema porque hacen un uso más eficiente de las partículas. Pero, por ejemplo, la evaluación de pesos en un filtro de partículas como el filtro bootstrap en altas dimensiones se vuelve problemática porque la evaluación de la verosimilitud da valores demasiado próximos a cero. Por estos motivos, la utilización del EM *online* en espacios de gran dimensionalidad constituye una línea de investigación interesante. En particular, resulta relevante estudiar el desempeño de la metodología en combinación con las estrategias clásicas de asimilación de datos para paliar algunos de estos problemas: por ejemplo, la inflación y la localización. Por otro lado se requiere de propuestas de parametrización de la covarianza ya que no es posible representarla explícitamente en espacios de alta dimensionalidad.

Otras posibles extensiones del EM *online* involucran la combinación de la metodología con filtros de Kalman por ensambles más sofisticados. Por simplicidad nosotros usamos el filtro estocástico (Burgers et al., 1998), sin embargo existe una familia vasta de filtros por ensambles con los que evaluar la técnica. De manera similar, a pesar de que usamos el algoritmo en combinación con un filtro de partículas moderno (el VMPF), la variedad de filtros de partículas es muy amplia y sería valioso estudiar cómo se desempeña el EM *online* en este espectro de métodos. Por otro lado, la implementación del algoritmo con un paso de suavizado hacia atrás sugiere la posibilidad de una implementación que realice un suavizado de varios pasos temporales hacia atrás. Esta variación involucraría una ventana móvil, de manera que el suavizado hacia atrás tiene una interpretación que se denomina *fixed-lagged smoother* (Cosme et al., 2012). Estas ventanas se superpondrían, tal como en una implementación de media móvil pero también existe la posibilidad de considerar ventanas yuxtapuestas en las que se corre el algoritmo EM *batch* clásico, es decir una suerte de implementación con *mini-batchs*. Una adaptación extra que vale la pena considerar es la utilización del EM *online* con distribuciones no Gaussianas. La metodología considera distribuciones de la familia exponencial y en el caso Gaussiano se tiene la ventaja de que la solución que anula al gradiente de la ELBO se puede computar de manera explícita pero esto puede no ser el caso para otras distribuciones.

La aplicación del EM *online* a modelos epidemiológicos compartimentales, en nuestro conocimiento, no se ha hecho antes. De hecho existen pocos antecedentes de estimación de errores en modelos epidemiológicos siendo una excepción el trabajo de Ionides et al., 2006 en el que se usan filtros iterados. Sin embargo, en este trabajo se aumenta el estado con parámetros estocásticos lo cual no garantiza buenas estimaciones (según DelSole y Yang, 2010). Los resultados muestran que la metodología del EM *online* es aplicable a modelos epidemiológicos compartimentales incluso incorporando el método de estado aumentado para parámetros del modelo con variación

temporal. Notamos que cuando se utiliza estado aumentado con caminatas aleatorias, el ruido que estas incorporan al sistema afecta sobre las estimaciones del error observacional y del sistema en general. En particular, si el ruido introducido es muy grande se puede llegar a un exceso de confianza sobre la calidad de las observaciones y por lo tanto a un sobreajuste. Esto se debe a que la estocasticidad de las caminatas aleatorias actúa como error de modelo y si este es muy grande, los pronósticos que produce el modelo pierden credibilidad. Para dar cuenta de esta situación existen algunas alternativas. Por un lado se pueden interpretar a las varianzas de las distribuciones de las caminatas como parte de una matriz $\mathbf{Q}$ que parametrice el error de todo el modelo y estimar con el EM *online* a esta matriz. Por otro lado es posible utilizar métodos de inflación de ensamble adaptativos (Ruiz et al., 2013b).

Otra observación que se debe hacer respecto a la aplicación del EnKF en sistemas epidemiológicos tiene que ver con que las variables son definidas positivas. Esto no es muy problemático cuando los valores que toman las variables están lejos del cero puesto que las incertezas Gaussianas pueden ser apropiadas. Sin embargo, cerca del cero las distribuciones Gaussianas pueden resultar inexactas y dar lugar a valores negativos sin interpretación epidemiológica en el sistema. Este problema se extiende también a la asimilación con EnKF en ABMs. Una posible solución al problema es la utilización de filtros de Kalman que admiten restricciones lineales, como que las variables sean positivas o que la suma de los compartimentos sea igual al total de la población (Gupta, 2007). Por otro lado también existen versiones del EnKF adaptadas para variables definidas positivas utilizando distribuciones Gamma o Gamma inversa (Bishop, 2016).

La utilización de asimilación de datos por ensambles en ABMs mostró ser factible en su aplicación a un modelo epidemiológico sencillo. Los experimentos muestran que la metodología permitió estimar correctamente las variables macroscópicas del sistema junto con parámetros epidemiológicos de interés, como la cantidad de contactos diarios o la proporción de casos no reportados. A pesar de que Ward et al., 2016 también plantea la combinación de ABMs con el EnKF, no se hace referencia en ese trabajo a como debería modificarse la población de agentes para ser consistente a las correcciones del EnKF sobre el estado macroscópico. Nosotros planteamos dos maneras de hacer esto para el caso particular de nuestro modelo epidemiológico pero creemos que algunas de las ideas subyacentes, sobre todo del método de redistribución aleatorizada, pueden generalizarse a una gama más amplia de ABMs. La utilización de metodologías de esta naturaleza sobre modelos más complejos es un problema que hasta donde conocemos no ha sido explorado pero que podría dar resultados interesantes. Esto se debe a que la inferencia sobre ABMs y su calibración es importante para que este tipo de modelos puedan mejorar su desempeño y para facilitar su uso en sistemas operacionales de predicción. En este sentido, los modelos dinámicos de ecuaciones diferenciales han sido mucho más estudiados por lo que existen actualmente muchas más herramientas de inferencia diseñadas y evaluadas para este tipo de modelo.

Nuestra metodología se basa en asimilar datos en el espacio de las variables macroscópicas que también son en general las variables observadas. Sin embargo es posible pensar en asimilar datos al nivel microscópico. Para esto podemos considerar a todos los atributos de los agentes como el estado oculto del modelo de Markov escondido pero esto trae algunos inconvenientes que deben ser resueltos. Por un lado, los atributos de los agentes pueden pertenecer a espacios en los que no se puede aplicar el EnKF de manera directa. Por ejemplo si los atributos son números enteros o etiquetas no queda claro como utilizar el EnKF porque este está diseñado para valores reales. Por otro lado, muchos de los atributos pueden no estar

correlacionados a las observaciones. Además de esto, la cantidad de atributos puede ser muy grande, mucho más que el de las variables agregadas y por otro lado, la distribución de los atributos no es necesariamente Gaussiana lo que potencialmente degrada el desempeño del EnKF. Asimilar datos a este nivel habilitaría también el uso de observaciones a nivel micro. De hecho, los datos individualizados son cada vez más comunes por el uso de dispositivos personales y el monitoreo de actividad en internet. Este enfoque permitiría una combinación potencialmente fructífera entre este tipo de datos y los ABMs.

Una de las características del esquema que propusimos para asimilar datos en ABMs es que las poblaciones de agentes se corrigen para ser consistentes con las variables del macroestado. Esta modificación en los atributos de agentes puede tener consecuencias difíciles de pronosticar. Una alternativa que evitaría estos cambios podría ser la de utilizar un filtro de partículas *bootstrap*. Este filtro, para construir la muestra del análisis, toma la muestra del pronóstico y remuestrea las partículas de acuerdo a cuán verosímil hacen a la observación. Esto significa que las poblaciones de agentes no necesitan ser modificadas en absoluto, simplemente se remuestrea el ensamble y se conservan como análisis las poblaciones de agentes seleccionadas, con las repeticiones correspondientes provenientes del remuestreo si las hubiera. Es necesario remarcar que el hecho de que el filtro *bootstrap* reutilice las partículas del pronóstico para construir el análisis, es conveniente para la aplicación con ABMs pero no es una característica deseable en general puesto que da lugar a la llamada pérdida de la diversidad. De hecho, filtros de partículas más modernos usan algún tipo de mutación o *jittering* para evitar que existan partículas repetidas y que se mantenga la diversidad para una mejor representación de las distribuciones. De utilizar algún tipo de mutación, en la aplicación con ABMs, se debería entonces intervenir sobre los atributos de los agentes. Concluímos que la combinación de ABMs con otras herramientas de las que dispone la asimilación de datos puede dar lugar interacciones fructíferas y vadría la pena explorar estas posibilidades.

Finalmente, resta investigar el efecto de los errores observacionales y de modelo en ABMs y en particular en sistemas combinados con asimilación de datos. Es posible pensar en la introducción de modelo aditivo o mediante inflación sobre las macrovariables y luego, mediante alguna corrección como las que propusimos en la Sección 5.3.1 dar cuenta de esos cambios sobre la población de agentes. Otra opción sería introducir error directamente sobre los atributos de los agentes aunque no queda claro que utilizar distribuciones Gaussianas sea adecuado en este caso. De hecho, sobre atributos que no son números reales se debe considerar otra alternativa. En cuanto al error observacional sobre las variables macroscópicas se tiene, al igual que en los modelos compartimentales, que las variables son definidas positivas. Como mencionamos antes esto puede causar problemas para el EnKF. El error en este tipo de variables tiende a depender de la magnitud de la variable en sí por lo que para estimarla sería adecuado utilizar algún método adaptativo como el EM *online*. La estimación de incertezas en ABMs es un área poco estudiada pero es un tema que deberá ser atendido para que los ABMs se consoliden como herramientas para el modelado e inferencia de sistemas complejos.

Apéndice A

Asimilación de datos

A.1 Algoritmo *forward-backward*

El algoritmo *forward-backward* está especificado en 10.

Algoritmo 10: Algoritmo forward filter backward smoothing

input :
Distribución inicial $p(\mathbf{x}_0)$
Distribución de transición $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ para $t = 1, ..., T$
Verosimilitud observacional $p(\mathbf{y}_t|\mathbf{x}_t)$ para $t = 1, ..., T$
output:
Distribución predictiva $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$ para para $t = 1, ..., T$
Distribución filtrante $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ para $t = 1, ..., T$
Distribución suavizante $p(\mathbf{x}_t|\mathbf{y}_{1:T})$ para $t = 1, ..., T$

Forward filter
for $t = 1, ..., T$ **do**
  Computar distribución predictiva:
  $p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})d\mathbf{x}_{t-1}$
  Computar distribución filtrante: $p(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$
end
Backward smoother
for $t = T, ..., 1$ **do**
  Computar distribución suavizante:
  $p(\mathbf{x}_t|\mathbf{y}_{1:T}) = \int \frac{p(\mathbf{x}_{t+1}|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t})}{p(\mathbf{x}_{t+1}|\mathbf{y}_{1:t})} p(\mathbf{x}_{t+1}|\mathbf{y}_{1:T})d\mathbf{x}_{t+1}$
end

La distribución predictiva se puede deducir integrando la distribución de transición pesando con la distribución filtrante del paso anterior:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) = \int p(\mathbf{x}_t, \mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})d\mathbf{x}_{t-1} \quad \text{Marginalización} \quad (A.1)$$

$$= \int p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_{1:t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})d\mathbf{x}_{t-1} \quad \text{Bayes} \quad (A.2)$$

$$= \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})d\mathbf{x}_{t-1} \quad \text{Propiedades HMM} \quad (A.3)$$

Por otro lado, para obtener la distribución filtrante podemos usar la distribución predictiva e incorporar la información de la observación $\mathbf{y}_t$ de la siguiente manera:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) = \frac{p(\mathbf{y}_t|\mathbf{x}_t\mathbf{y}_{1:t-1})p(\mathbf{x}_t|\mathbf{y}_{1:t-1})}{p(\mathbf{y}_t|\mathbf{y}_{1:t-1})} \qquad \text{Bayes} \tag{A.4}$$

$$= \frac{p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1})}{p(\mathbf{y}_t|\mathbf{y}_{1:t-1})} \qquad \text{Propiedades HMM} \tag{A.5}$$

$$\propto p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) \tag{A.6}$$

Para calcular la distribución suavizante necesitamos tener las distribuciones filtrantes y predictivas del forward-pass e iterar desde la última observación hasta la primera:

$$p(\mathbf{x}_t|\mathbf{y}_{1:T}) = \int p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:T})p(\mathbf{x}_{t+1}|\mathbf{y}_{1:T})d\mathbf{x}_{t+1} \qquad \text{Marginalización} \tag{A.7}$$

$$= \int p(\mathbf{x}_t|\mathbf{x}_{t+1}, \mathbf{y}_{1:t})p(\mathbf{x}_{t+1}|\mathbf{y}_{1:T})d\mathbf{x}_{t+1} \qquad \text{Propiedades HMM} \tag{A.8}$$

$$= \int \frac{p(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{y}_{1:t})p(\mathbf{x}_t|\mathbf{y}_{1:t})}{p(\mathbf{x}_{t+1}|\mathbf{y}_{1:t})} p(\mathbf{x}_{t+1}|\mathbf{y}_{1:T})d\mathbf{x}_{t+1} \qquad \text{Bayes} \tag{A.9}$$

$$= \int \frac{p(\mathbf{x}_{t+1}|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t})}{p(\mathbf{x}_{t+1}|\mathbf{y}_{1:t})} p(\mathbf{x}_{t+1}|\mathbf{y}_{1:T})d\mathbf{x}_{t+1} \qquad \text{Propiedades HMM} \tag{A.10}$$

A.2 Filtro de Kalman

La fórmula 2.12 para la media $\mathbf{x}_t^f$ de la distribución predictiva puede ser deducida de la siguiente manera:

$$\begin{aligned}
\mathbf{x}_t^f &= E[\mathbf{X}_t|\mathbf{y}_{1:t-1}] && \text{Definición de } \mathbf{x}_t^f \\
&= \int \mathbf{x}_t p(\mathbf{x}_t|y_{1:t-1})d\mathbf{x}_t && \text{Definición de } E \\
&= \int \mathbf{x}_t \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})d\mathbf{x}_{t-1}d\mathbf{x}_t && \text{Ec. 2.7} \\
&= \int p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) \int \mathbf{x}_t p(\mathbf{x}_t|\mathbf{x}_{t-1})d\mathbf{x}_t d\mathbf{x}_{t-1} && \text{Intercambio de } \int \\
&= \int p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})E[\mathbf{X}_t|\mathbf{x}_{t-1}]d\mathbf{x}_{t-1} && \text{Definición de } E \\
&= \int p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})E[\mathbf{M}_t\mathbf{x}_{t-1} + \boldsymbol{\eta}_t]d\mathbf{x}_{t-1} && \text{Modelo} \\
&= \int p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})\mathbf{M}_t\mathbf{x}_{t-1}d\mathbf{x}_{t-1} && \mathbf{M}_t \text{ lineal y } E[\boldsymbol{\eta}_t] = 0 \\
&= \mathbf{M}_t \int p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})\mathbf{x}_{t-1}d\mathbf{x}_{t-1} && \text{Modelo} \\
&= \mathbf{M}_t E[\mathbf{X}_{t-1}|\mathbf{y}_{1:t-1}] && \mathbf{M}_t \text{ lineal} \\
&= \mathbf{M}_t\mathbf{x}_{t-1}^a && \text{Definición de } \mathbf{x}_t^a
\end{aligned}$$

Por otro lado, la fórmula 2.13 para la matriz de covarianza $\mathbf{P}_t^f$ de la distribución predictiva puede ser obtenida como se detalla a continuación:

$$\begin{aligned}
\mathbf{P}_t^f &= Var[\mathbf{X}_t|\mathbf{y}_{1:t-1}] && \text{Definición de } \mathbf{P}_t^f \\
&= E[\mathbf{X}_t\mathbf{X}_t^T|\mathbf{y}_{1:t-1}] - E[\mathbf{X}_t|\mathbf{y}_{1:t-1}]E[\mathbf{X}_t|\mathbf{y}_{1:t-1}]^T && Var[\mathbf{X}] = E[\mathbf{X}\mathbf{X}^T] - E[\mathbf{X}]E[\mathbf{X}]^T \\
&= E[\mathbf{X}_t\mathbf{X}_t^T|\mathbf{y}_{1:t-1}] - \mathbf{M}_t\mathbf{x}_{t-1}^a\mathbf{x}_{t-1}^{aT}\mathbf{M}_t^T && \text{Ec. 2.12}
\end{aligned}$$

Ahora desarrollamos el valor esperado del primer término:

$$\begin{aligned}
E[\mathbf{X}_t\mathbf{X}_t^T|\mathbf{y}_{1:t-1}] &= \int \mathbf{x}_t\mathbf{x}_t^T p(\mathbf{x}_t|\mathbf{y}_{1:t-1})d\mathbf{x}_t \\
&= \int \mathbf{x}_t\mathbf{x}_t^T \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})d\mathbf{x}_{t-1}d\mathbf{x}_t && \text{Ec. 2.7} \\
&= \int p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) \int \mathbf{x}_t\mathbf{x}_t^T p(\mathbf{x}_t|\mathbf{x}_{t-1})d\mathbf{x}_t d\mathbf{x}_{t-1} && \text{Intercambio de } \int \\
&= \int p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})E[\mathbf{X}_t\mathbf{X}_t^T|\mathbf{x}_{t-1}]d\mathbf{x}_{t-1} && \text{Definición de } E \\
&= \int p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})(Var[\mathbf{X}_t|\mathbf{x}_{t-1}] && E[\mathbf{X}\mathbf{X}^T] = Var[\mathbf{X}]^T + E[\mathbf{X}]E[\mathbf{X}]^T \\
&\quad + E[\mathbf{X}_t|\mathbf{x}_{t-1}]E[\mathbf{X}_t|\mathbf{x}_{t-1}]^T)d\mathbf{x}_{t-1} \\
&= \int p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})Var[\mathbf{X}_t|\mathbf{x}_{t-1}]d\mathbf{x}_{t-1} && \text{Linealidad de } \int \\
&\quad + \int p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})E[\mathbf{X}_t|\mathbf{x}_{t-1}]E[\mathbf{X}_t|\mathbf{x}_{t-1}]^T)d\mathbf{x}_{t-1} \\
&= \int p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})\mathbf{Q}_t d\mathbf{x}_{t-1} && \text{Ec. 2.10} \\
&\quad + \int p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})\mathbf{M}_t\mathbf{x}_{t-1}\mathbf{x}_{t-1}^T\mathbf{M}_t^T d\mathbf{x}_{t-1} \\
&= \mathbf{Q}_t + \mathbf{M}_t \int p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})\mathbf{x}_{t-1}\mathbf{x}_{t-1}^T d\mathbf{x}_{t-1}\mathbf{M}_t^T && \mathbf{M}_t \text{ lineal} \\
&= \mathbf{Q}_t + \mathbf{M}_t E[\mathbf{X}_{t-1}\mathbf{X}_{t-1}^T|\mathbf{y}_{1:t-1}]\mathbf{M}_t^T && \text{Definición de } E \\
&= \mathbf{Q}_t + \mathbf{M}_t E[\mathbf{X}_{t-1}|\mathbf{y}_{1:t-1}]E[\mathbf{X}_{t-1}|\mathbf{y}_{1:t-1}]^T\mathbf{M}_t^T && E[\mathbf{X}\mathbf{X}^T] = Var[\mathbf{X}] + E[\mathbf{X}]E[\mathbf{X}]^T \\
&\quad + \mathbf{M}_t Var[\mathbf{X}_{t-1}|\mathbf{y}_{1:t-1}]\mathbf{M}_t^T \\
&= \mathbf{Q}_t + \mathbf{M}_t\mathbf{x}_{t-1}^a\mathbf{x}_{t-1}^{aT}\mathbf{M}_t^T + \mathbf{M}_t\mathbf{P}_{t-1}^a\mathbf{M}_t^T && \text{Definición de } \mathbf{x}_{t-1}^a \text{ y } \mathbf{P}_{t-1}^a
\end{aligned}$$

Por lo tanto al combinar las expresiones obtenemos el resultado:

$$\mathbf{P}_t^f = \mathbf{Q}_t + \mathbf{M}_t\mathbf{P}_{t-1}^a\mathbf{M}_t^T$$

Para obtener las fórmulas de la media y covarianza de la distribución filtrante debemos usar la ecuación de análisis del algoritmo *forward-backwards*:

$$
\begin{aligned}
p(\mathbf{x}_t|\mathbf{y}_{1:t}) &\propto p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1}) && \text{Ec. 2.8}\\
&\propto \exp((\mathbf{y}_t - \mathbf{H}_t\mathbf{x}_t)^T\mathbf{R}_t^{-1}(\mathbf{y}_t - \mathbf{H}_t\mathbf{x}_t) && \text{Densidades Gaussianas}\\
&+ (\mathbf{x}_t - \mathbf{x}_t^f)^T(\mathbf{P}_t^f)^{-1}(\mathbf{x}_t - \mathbf{x}_t^f))\\
&\propto \exp(\mathbf{x}_t^T((\mathbf{P}_t^f)^{-1} + \mathbf{H}^T\mathbf{R}_t^{-1}\mathbf{H}_t)\mathbf{x}_t && \text{Distribución}\\
&- 2\mathbf{x}_t^T(\mathbf{H}^T\mathbf{R}_t^{-1}\mathbf{y}_t + (\mathbf{P}_t^f)^{-1}\mathbf{x}_t))\\
&= \exp(\mathbf{x}_t^T\mathbf{A}\mathbf{x}_t - 2\mathbf{x}_t^T\mathbf{v}) && \text{Renombre}\\
&= \exp((\mathbf{x}_t - \mathbf{A}^{-1}\mathbf{v})^T\mathbf{A}(\mathbf{x}_t - \mathbf{A}^{-1}\mathbf{v}) - \mathbf{v}^T\mathbf{A}\mathbf{v}) && \text{Completar cuadrados}\\
&\propto \exp((\mathbf{x}_t - \mathbf{A}^{-1}\mathbf{v})^T\mathbf{A}(\mathbf{x}_t - \mathbf{A}^{-1}\mathbf{v}))
\end{aligned}
$$

donde hemos utilizado la siguiente nomenclatura:

$$
\begin{aligned}
\mathbf{A} &= (\mathbf{P}_t^f)^{-1} + \mathbf{H}^T\mathbf{R}_t^{-1}\mathbf{H}_t\\
\mathbf{v} &= \mathbf{H}^T\mathbf{R}_t^{-1}\mathbf{y}_t + (\mathbf{P}_t^f)^{-1}\mathbf{x}_t
\end{aligned}
$$

.

La expresión que obtuvimos implica que la distribución filtrante es Gaussiana con media $\mathbf{A}^{-1}\mathbf{v}$ y covarianza $\mathbf{A}^{-1}$. Vamos a desarrollar estas expresiones para obtener la formulación clásica del filtro de Kalman. Para ello, necesitaremos usar la siguiente identidad matricial de Woodbury (Golub y Van Loan, 1996):

$$(\mathbf{A} + \mathbf{C}\mathbf{B}\mathbf{C}^T)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{C}(\mathbf{B}^{-1} + \mathbf{C}^T\mathbf{A}^{-1}\mathbf{C})^{-1}\mathbf{C}^T\mathbf{A}^{-1}$$

Tenemos entonces que:

$$
\begin{aligned}
\mathbf{P}_t^a &= \mathbf{A}^{-1}\\
&= ((\mathbf{P}_t^f)^{-1} + \mathbf{H}^T\mathbf{R}_t^{-1}\mathbf{H}_t)^{-1}\\
&= \mathbf{P}_t^f - \mathbf{P}_t^f\mathbf{H}_t^T(\mathbf{R}_t + \mathbf{H}\mathbf{P}_t^f\mathbf{H}^T)^{-1}\mathbf{H}_t\mathbf{P}_t^f && \text{Identidad de Woodbury}\\
&= \mathbf{P}_t^f - \mathbf{K}_t\mathbf{H}_t\mathbf{P}_t^f\\
&= (\mathbf{I} - \mathbf{K}_t\mathbf{H}_t)\mathbf{P}_t^f
\end{aligned}
$$

donde hemos definido a $\mathbf{K}_t = \mathbf{P}_t^f\mathbf{H}_t^T(\mathbf{R}_t + \mathbf{H}\mathbf{P}_t^f\mathbf{H}^T)^{-1}$. Esta matriz es denominada matriz de ganancia de Kalman. Para desarrollar la expresión de la media de la distribución además usaremos la notación $\mathbf{S}_t = (\mathbf{R}_t + \mathbf{H}\mathbf{P}_t^f\mathbf{H}^T)^{-1}$, con la cual la ganancia de Kalman se puede expresar como $\mathbf{K}_t = \mathbf{P}_t^f\mathbf{H}_t^T\mathbf{S}_t$ y podemos obtener la fórmula

para $\mathbf{x}_t^a$ de la siguiente manera:

$$\begin{aligned}
\mathbf{x}_t^a &= \mathbf{A}^{-1}\mathbf{v} \\
&= (\mathbf{I} - \mathbf{K}_t\mathbf{H}_t)(\mathbf{P}_t^f\mathbf{H}^T\mathbf{R}_t^{-1}\mathbf{y}_t + (\mathbf{P}_t^f)^{-1}\mathbf{x}_t) \\
&= \mathbf{x}_t - \mathbf{K}_t\mathbf{H}_t\mathbf{x}_t + \mathbf{P}_t\mathbf{H}_t^T\mathbf{R}_t^{-1}\mathbf{y}_t - \mathbf{K}_t\mathbf{H}_t\mathbf{P}_t\mathbf{H}_t^T\mathbf{R}_t^{-1}\mathbf{y}_t \\
&= \mathbf{x}_t - \mathbf{K}_t\mathbf{H}_t\mathbf{x}_t + \mathbf{P}_t\mathbf{H}_t^T\mathbf{R}_t^{-1}\mathbf{y}_t - \mathbf{K}_t\mathbf{H}_t\mathbf{P}_t\mathbf{H}_t^T\mathbf{R}_t^{-1}\mathbf{y}_t \\
&= \mathbf{x}_t - \mathbf{K}_t\mathbf{H}_t\mathbf{x}_t + \mathbf{P}_t\mathbf{H}_t^T\mathbf{S}_t\mathbf{S}_t^{-1}\mathbf{R}_t^{-1}\mathbf{y}_t - \mathbf{K}_t\mathbf{H}_t\mathbf{P}_t\mathbf{H}_t^T\mathbf{R}_t^{-1}\mathbf{y}_t \\
&= \mathbf{x}_t - \mathbf{K}_t\mathbf{H}_t\mathbf{x}_t + \mathbf{K}_t(\mathbf{R}_t + \mathbf{H}\mathbf{P}_t^f\mathbf{H}^T)^{-1}\mathbf{R}_t^{-1}\mathbf{y}_t - \mathbf{K}_t\mathbf{H}_t\mathbf{P}_t\mathbf{H}_t^T\mathbf{R}_t^{-1}\mathbf{y}_t \\
&= \mathbf{x}_t - \mathbf{K}_t\mathbf{H}_t\mathbf{x}_t + \mathbf{K}_t\mathbf{y}_t \\
&= \mathbf{x}_t + \mathbf{K}_t(\mathbf{y}_t - \mathbf{H}_t\mathbf{x}_t)
\end{aligned}$$

A.3 Filtro de partículas

Hacemos aquí una deducción de el filtro de partículas SIR. El objetivo es obtener representaciones de partículas $\{\mathbf{x}_t^{(i)}, w_t^{(i)}\}_{i=1}^N$ tal que sea una aproximación empírica de $p(\mathbf{x}_t|\mathbf{y}_{1:t})$. Vamos a comenzar por considerar la distribución conjunta de las variables de estado condicionadas a las observaciones, $p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$. Estamos considerando a las variables de estado desde el tiempo 0 hasta el t, lo cual significa que esta es la densidad de probabilidad de una trayectoria de las variables de estado en el tiempo. Claramente, si tenemos una muestra de esta distribución, las componentes correspondientes al tiempo t constituirán una muestra de la probabilidad filtrante marginalizada $p(\mathbf{x}_t|\mathbf{y}_{1:t})$. Utilizando las propiedades Markovianas y la independencia condicional de las observaciones del modelo de Markov escondido podemos escribir:

$$p(\mathbf{x}_{0:t}|\mathbf{y}_{1:t}) \propto p(\mathbf{x}_{0:t-1}|\mathbf{y}_{1:t})p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{y}_t|\mathbf{x}_t)$$

Si muestreamos trayectorias $\{\mathbf{x}_{0:t}^{(i)}\}_{i=1}^N$ de una probabilidad propuesta q vamos a obtener que los pesos de importancia son

$$w_t^{(i)} \propto \frac{p(\mathbf{x}_{0:t-1}|\mathbf{y}_{1:t})p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{y}_t|\mathbf{x}_t)}{q(\mathbf{x}_{0:t})} \tag{A.11}$$

Adicionalmente consideraremos que q cumple

$$q(\mathbf{x}_{0:t}|\mathbf{y}_{1:t}) = q(\mathbf{x}_t|\mathbf{x}_{0:t-1}, \mathbf{y}_{1:t})q(\mathbf{x}_{0:t-1}|\mathbf{y}_{1:t-1}) \tag{A.12}$$

Esta factorización implica que si tenemos una muestra de la trayectoria $\mathbf{x}_{0:t-1}^{(i)} \sim q(\mathbf{x}_{0:t-1}|\mathbf{y}_{1:t-1})$ entonces se puede obtener una muestra de la trayectoria hasta el tiempo t incorporando la última componente muestreada como $\mathbf{x}_t^{(i)} \sim q(\mathbf{x}_t|\mathbf{x}_{0:t-1}, \mathbf{y}_{1:t})$.

Si entonces introducimos A.12 en A.11, tenemos que los pesos de importancia pueden ser computados como:

$$\begin{aligned}
w_t^{(i)} &\propto \frac{p(\mathbf{x}_{0:t-1}|\mathbf{y}_{1:t})p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{y}_t|\mathbf{x}_t)}{q(\mathbf{x}_{0:t-1}|\mathbf{y}_{1:t-1})q(\mathbf{x}_t|\mathbf{x}_{0:t-1}, \mathbf{y}_{1:t})} \\
&\propto w_{t-1}^{(i)} \frac{p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{y}_t|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{0:t-1}, \mathbf{y}_{1:t})}
\end{aligned}$$

Si adicionalmente dotamos a q de "Markovianidad" en el sentido que $q(\mathbf{x}_t|\mathbf{x}_{0:t-1}\mathbf{y}_{1:t}) = q(\mathbf{x}_t|\mathbf{x}_{t-1}\mathbf{y}_t)$, entonces los pesos solamente dependen de $\mathbf{x}_t$ y no de toda la trayectoria $\mathbf{x}_{0:t-1}$. De esta manera se puede hacer filtrado de manera secuencial. Con estas suposiciones obtenemos la forma general de los pesos de el filtro de partículas SIR:

$$w_t^{(i)} \propto w_{t-1}^{(i)} \frac{p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{y}_t|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{y}_t)}$$

A.4 EnKF

Para mostrar que la formulación del EnKF estocástico es correcta, basta probar que la media y covarianza que se obtiene de considerar a los miembros del ensamble como variables aleatorias coinciden con las del filtro de Kalman tradicional, el cual da una solución exacta al problema. Supongamos inductivamente que las partículas a tiempo $t-1$ cumplen con esto y veamos que podemos obtener las fórmulas correctas a tiempo t.

Las partículas del pronóstico están definidas como

$$\mathbf{x}_t^{f,(i)} = \mathbf{M}_t \mathbf{x}_{t-1}^{a,(i)} + \boldsymbol{\eta}_t^{(i)}$$

con $\boldsymbol{\eta}_t^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_t)$. Por lo tanto si calculamos la media y covarianza, obtenemos las mismas fórmulas que para el filtro de Kalman tradicional 2.12 y 2.13:

$$\begin{aligned}
E[\mathbf{x}_t^{f,(i)}] &= E[\mathbf{M}_t \mathbf{x}_{t-1}^{a,(i)} + \boldsymbol{\eta}_t^{(i)}] \\
&= \mathbf{M}_t E[\mathbf{x}_{t-1}^{a,(i)}] + E[\boldsymbol{\eta}_t^{(i)}] \\
&= \mathbf{M}_t \mathbf{x}_{t-1}^{a} \\
Var[\mathbf{x}_t^{f,(i)}] &= Var[\mathbf{M}_t \mathbf{x}_{t-1}^{a,(i)} + \boldsymbol{\eta}_t^{(i)}] \\
&= \mathbf{M}_t Var[\mathbf{x}_{t-1}^{a,(i)}] \mathbf{M}_t^T + Var[\boldsymbol{\eta}_t^{(i)}] \\
&= \mathbf{M}_t \mathbf{P}_t^a \mathbf{M}_t^T + \mathbf{Q}_t
\end{aligned}$$

donde hemos usado la independencia de $\boldsymbol{\eta}_t^{(i)}$ y $\mathbf{x}_{t-1}^{a,(i)}$ y que $Var[\mathbf{AX}] = \mathbf{A}Var[\mathbf{X}]\mathbf{A}^T$ cuando $\mathbf{A}$ es un operador lineal. Las partículas del pronóstico entonces constituyen efectivamente una muestra de la distribución del pronóstico.

Las partículas del análisis en el EnKF estocástico están definidas por

$$\mathbf{x}_t^{a,(i)} = \mathbf{x}_t^{f,(i)} + \widehat{\mathbf{K}_t}(\mathbf{y}_t - (\mathbf{H}_t \mathbf{x}_t^{f,(i)} + \mathbf{v}_t^{(i)}))$$

donde $\mathbf{v}_t^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_t)$ y la aproximación a la ganancia de Kalman es $\widehat{\mathbf{K}}_t = \widehat{\mathbf{P}}_t^f \mathbf{H}_t^T (\mathbf{R}_t + \mathbf{H}_t \widehat{\mathbf{P}}_t^f \mathbf{H}_t^T)^{-1}$ con $\widehat{\mathbf{P}}_t^f$ denotando la covarianza muestral del ensamble de pronóstico. Probaremos que en el caso ideal en que estemos usando la matriz de ganancia verdadera (es decir la que utiliza la covarianza exacta del pronóstico, $\mathbf{P}_t^f$) la media y covarianza del ensamble del análisis coinciden con las del filtro de Kalman 2.14 y

2.15.

$$\begin{aligned}
E[\mathbf{x}_t^{a,(i)}] &= E[\mathbf{x}_t^{f,(i)} + \mathbf{K}_t(\mathbf{y}_t - (\mathbf{H}_t\mathbf{x}_t^{f,(i)} + \mathbf{v}_t^{(i)}))] \\
&= \mathbf{x}_t^f + \mathbf{K}_t E[(\mathbf{y}_t - (\mathbf{H}_t\mathbf{x}_t^{f,(i)} + \mathbf{v}_t^{(i)}))] \\
&= \mathbf{x}_t^f + \mathbf{K}_t(\mathbf{y}_t - E[\mathbf{H}_t\mathbf{x}_t^{f,(i)}]) \\
&= \mathbf{x}_t^f + \mathbf{K}_t(\mathbf{y}_t - \mathbf{H}_t\mathbf{x}_t^f)
\end{aligned}$$

Por otro lado, para la varianza del ensamble de análisis tenemos que:

$$\begin{aligned}
Var[\mathbf{x}_t^{a,(i)}] &= Var[\mathbf{x}_t^{f,(i)} + \mathbf{K}_t(\mathbf{y}_t - (\mathbf{H}_t\mathbf{x}_t^{f,(i)} + \mathbf{v}_t^{(i)}))] \\
&= Var[\mathbf{x}_t^{f,(i)}] + Var[\mathbf{K}_t(\mathbf{y}_t - (\mathbf{H}_t\mathbf{x}_t^{f,(i)} + \mathbf{v}_t^{(i)}))] \\
&+ 2Cov[\mathbf{x}_t^{f,(i)}, \mathbf{K}_t(\mathbf{y}_t - (\mathbf{H}_t\mathbf{x}_t^{f,(i)} + \mathbf{v}_t^{(i)}))] \\
&= \mathbf{P}_t^f + \mathbf{K}_t Var[(\mathbf{H}_t\mathbf{x}_t^{f,(i)} + \mathbf{v}_t^{(i)})]\mathbf{K}_t^T + 2Cov[\mathbf{x}_t^{f,(i)}, -\mathbf{K}_t\mathbf{H}_t\mathbf{x}_t^{f,(i)}] \\
&= \mathbf{P}_t^f + \mathbf{K}_t(\mathbf{H}_t\mathbf{P}_t^f\mathbf{H}_t^T + \mathbf{R}_t)\mathbf{K}_t^T - 2\mathbf{K}_t\mathbf{H}_t\mathbf{P}_t^f \\
&= \mathbf{P}_t^f + \mathbf{P}_t^f\mathbf{H}_t^T(\mathbf{H}_t\mathbf{P}_t^f\mathbf{H}_t^T + \mathbf{R}_t)^{-1}(\mathbf{H}_t\mathbf{P}_t^f\mathbf{H}_t^T + \mathbf{R}_t)\mathbf{K}_t^T - 2\mathbf{K}_t\mathbf{H}_t\mathbf{P}_t^f \\
&= \mathbf{P}_t^f + \mathbf{P}_t^f\mathbf{H}_t^T\mathbf{K}_t^T - 2\mathbf{K}_t\mathbf{H}_t\mathbf{P}_t^f \\
&= \mathbf{P}_t^f + (\mathbf{K}_t\mathbf{H}_t\mathbf{P}_t^f)^T - 2\mathbf{K}_t\mathbf{H}_t\mathbf{P}_t^f \\
&= \mathbf{P}_t^f - \mathbf{K}_t\mathbf{H}_t\mathbf{P}_t^f \\
&= (\mathbf{I} - \mathbf{K}_t\mathbf{H}_t)\mathbf{P}_t^f
\end{aligned}$$

donde hemos usado la propiedad de la varianza de una suma ($Var[\mathbf{X}+\mathbf{Y}] = Var[\mathbf{X}] + Var[\mathbf{Y}] + Cov[\mathbf{X},\mathbf{Y}]$), la independencia de $\mathbf{x}_t^{f,(i)}$ con $\mathbf{v}_t^{(i)}$ y el hecho que $\mathbf{K}_t\mathbf{H}_t\mathbf{P}_t^f$ es una matriz simétrica.

Apéndice B

Algoritmo EM

B.1 Gaussiana multivariada como miembro de la familia exponencial

Aquí presentamos la representación de densidades de probabilidad de la familia exponencial y cómo se puede escribir a la Gaussiana como miembro de dicha familia. Luego damos la solución de la Ecuación 3.8 para encontrar el valor del parámetro que anula a la ELBO en el caso Gaussiano.

Se dice que una probabilidad pertenece a la familia exponencial si su densidad de probabilidad, parametrizada por $\boldsymbol{\theta}$ puede ser escrita como:

$$p(\mathbf{x};\boldsymbol{\theta}) = h(\mathbf{x})\exp(\psi(\boldsymbol{\theta})\cdot S(\mathbf{x}) - A(\boldsymbol{\theta}))$$

donde $S(\mathbf{x})$ es llamado el estadístico suficiente, $\psi(\boldsymbol{\theta})$ la parametrización natural y h y A son funciones bien definidas (Wasserman, 2004). Es importante en esta representación, que la interacción entre $\mathbf{x}$ y $\boldsymbol{\theta}$ se produce solamente a través del producto interno $\psi(\boldsymbol{\theta})\cdot S(\mathbf{x})$.

La densidad de probabilidad de una Gaussiana multivariada de dimensionalidad N con media $\boldsymbol{\mu}$ y covarianza $\boldsymbol{\Sigma}$ es habitualmente expresada como:

$$p(\mathbf{x};\boldsymbol{\theta}) = (2\pi)^{-\frac{k}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}$$

Nuestro objetivo es expresar esta función con la forma de la familia exponencial considerando como parámetro solamente a la varianza $\boldsymbol{\Sigma}$ y considerando que $\boldsymbol{\mu}$ es conocido. Para ello consideremos la extensión del producto punto entre vectores a matrices como el producto Frobenious (elemento a elemento). Esto permite que la forma cuadrática dentro de la exponencial de la densidad Gaussiana se pueda reescribir como:

$$(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}\cdot((\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T)$$

y por lo tanto logramos la forma de la familia exponencial usando:

$$\begin{aligned} h(\mathbf{x}) &= (2\pi)^{-\frac{N}{2}} \\ \psi(\boldsymbol{\theta}) &= -\frac{1}{2}\boldsymbol{\Sigma}^{-1} \\ S(\mathbf{x}) &= (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T \\ A(\boldsymbol{\theta}) &= \frac{1}{2}\log|\boldsymbol{\Sigma}|. \end{aligned}$$

B.2 Punto crítico de la ELBO en caso Gaussiano

Aquí damos una solución para la Ecuación 3.8 en el caso Gaussiano utilizando la expresión de la densidad de probabilidad como miembro de la familia exponencial desarrollada en B.1:

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta})\cdot S - \nabla_{\boldsymbol{\theta}}A(\boldsymbol{\theta}) &= 0 \\
-\frac{1}{2}\nabla_{\boldsymbol{\Sigma}}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \nabla_{\boldsymbol{\Sigma}}\frac{1}{2}\log|\boldsymbol{\Sigma}| &= 0 \\
\frac{1}{2}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1} - \frac{1}{2}\boldsymbol{\Sigma}^{-1} &= 0 \\
\boldsymbol{\Sigma}^{-1}S\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1} &= 0 \\
S &= \boldsymbol{\Sigma}
\end{aligned}$$

donde hemos usado la expresión $\frac{\partial \log|\mathbf{X}|}{\partial \mathbf{X}} = \mathbf{X}^{-T}$ para la derivada del logaritmo del determinante y $\frac{\partial \mathbf{a}^T\mathbf{X}^{-1}\mathbf{b}}{\partial \mathbf{X}} = \mathbf{X}^{-T}\mathbf{a}\mathbf{b}^T\mathbf{X}^{-1}$ para la derivada de la forma cuadrática Petersen y Pedersen, 2012.

B.3 Factorización de $p(\mathbf{x}_{t-1}, \mathbf{x}_t|\mathbf{y}_{1:t})$

Desarrollamos aquí la factorización de la probabilidad conjunta utilizada en 3.22.

$$\begin{aligned}
p(\mathbf{x}_{t-1},\mathbf{x}_t|\mathbf{y}_{1:t}) &= p(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{y}_{1:t})p(\mathbf{x}_t|\mathbf{y}_{1:t}) && \text{Bayes} && \text{(B.1)}\\
&= p(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{y}_{1:t-1})p(\mathbf{x}_t|\mathbf{y}_{1:t}) && \text{Prop. HMM} && \text{(B.2)}\\
&= \frac{p(\mathbf{x}_t|\mathbf{x}_{t-1},\mathbf{y}_{1:t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})}{p(\mathbf{x}_t|\mathbf{y}_{1:t-1})}p(\mathbf{x}_t|\mathbf{y}_{1:t}) && \text{Bayes} && \text{(B.3)}\\
&= p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})\frac{p(\mathbf{x}_t|\mathbf{y}_{1:t})}{p(\mathbf{x}_t|\mathbf{y}_{1:t-1})} && \text{Prop. HMM} && \text{(B.4)}\\
&= p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})\frac{p(\mathbf{y}_t|\mathbf{x}_t)}{p(\mathbf{y}_t|\mathbf{y}_{1:t-1})} && \text{Ecuación A.5} && \text{(B.5)}
\end{aligned}$$

B.4 Aproximación de la verosimilitud

Desarrollamos aquí la aproximación de Monte Carlo utilizada en 3.14.

$$\begin{aligned}
\log p(\mathbf{y}_{1:T}) &= \log\prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) \\
&= \log\prod_{t=1}^{T}\int p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1})d\mathbf{x}_t \\
&= \sum_{t=1}^{T}\log\int p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}_{1:t-1})d\mathbf{x}_t \\
&\approx \sum_{t=1}^{T}\log\frac{1}{N_p}\sum_{j=1}^{N_p} p(\mathbf{y}_t|\mathbf{x}_t^{f,(j)})
\end{aligned}$$

donde las partículas $\{\mathbf{x}_t^{f,(j)}\}_{j=1}^{N_p}$ están muestreadas de la distribución de pronóstico $p(\mathbf{x}_t|\mathbf{y}_{1:t-1})$. La identidad utilizada en la primera línea se puede encontrar en Carrassi, Bocquet, Hannart et al., 2017

Apéndice C

Modelo epiABM

C.1 Parametrización por defecto

Aquí especificamos la parametrización por defecto que utilizamos para el modelo epiABM. El Cuadro C.1 sintetiza estos valores que fueron elegidos para describir la etapa inicial de la pandemia de COVID-19. Como el modelo se utiliza solamente para evaluar la metodología de inferencia no tenemos pretensión de dar valores equivalentes a los de estudios médicos sino que solamente tomamos valores razonables para nuestros objetivos. En Guan et al., 2020 se reporta un período de incubación de 4 días el cual es consistente con nuestra parametrización de la distribución Gamma ($\mu_E = k_E\theta_E = 4$ días). El tiempo medio de la etapa infecciosa elegido para casos leves es $\mu_{I_M} = k_{I_M}\theta_{I_M} = 8$ días y valores similares fueron utilizados en otros modelos (Ivorra et al., 2020; Zhao y Feng, 2020). Para el tiempo entre la enfermedad grave y a hospitalización usamos $\mu_{I_S} = k_{I_S}\theta_{I_S} = 8.1$ días lo cual está en el rango reportado en Faes et al., 2020. Tomamos a la probabilidad de desarrollar sintomatología grave, q_S, como 10% y a la probabilidad de que un paciente hospitalizado muera como 40%. Esto resulta en una fatalidad total de 4% que es alta pero en línea con los valores iniciales de la pandemia: en China, en Febrero del 2020 se registró un valor de 3.67% (Verity et al., 2020). En Argentina se encontraron valores similares en los experimentos preliminares publicados en Evensen et al., 2020. Este valor disminuyó sustancialmente pasada la primera etapa de la pandemia y aún más con el el comienzo de las vacunaciones masivas y el surgimiento de variantes menos letales. Para el valor de λ, que parametriza el valor medio de contactos diarios de cada agente, utilizamos distintas configuraciones en todos los experimentos por lo cual no proveemos un valor por defecto. La cantidad esperada de infecciones que un agente infectado produce en una población totalmente susceptible se puede computar como $\lambda\beta$. Con las elecciones que hicimos para λ y los valores por defecto para β_C, β_D y q_C tenemos que esta cantidad es similar a la que se determina en los valores por defecto en el modelo en Kerr et al., 2020. El vector de probabilidades para la distribución del tamaño de hogares por defecto lo tomamos como $p_H = (0.36 \quad 0.27 \quad 0.16 \quad 0.13 \quad 0.08)$. Esto implica que solo consideramos casas de hasta 5 habitantes. Estos valores son basados en la Encuesta Anual de Hogares 2019 para la Ciudad Autónoma de Buenos Aires.

Tabla C.1: Parametrización por defecto para el epiABM.

Parámetro	Descripción	Valor
β_D	Probabilidad de infección en contactos domésticos	0.8
β_C	Probabilidad de infección en contactos casuales	0.16
q_D	Probabilidad de muerte para hospitalizados	0.4
q_S	Probabilidad de que una infección sea grave	0.1
q_C	Probabilidad de que un contacto sea casual	0.5
k_E	Parámetro de forma para la Gamma correspondiente a E	1.78
θ_E	Parámetro de escala para la Gamma correspondiente a E	2.25
k_{I_M}	Parámetro de forma para la Gamma correspondiente a I_M	7.11
θ_{I_M}	Parámetro de escala para la Gamma correspondiente a I_M	1.13
k_{I_S}	Parámetro de forma para la Gamma correspondiente a I_S	4.0
θ_{I_S}	Parámetro de escala para la Gamma correspondiente a I_S	1.0
k_H	Parámetro de forma para la Gamma correspondiente a H	9.0
θ_H	Parámetro de escala para la Gamma correspondiente a H	0.9

Bibliografía

Aanonsen, S. I., Nævdal, G., Oliver, D. S., Reynolds, A. C., & Vallès, B. (2009). The Ensemble Kalman Filter in Reservoir Engineering–a Review. *SPE Journal*, *14*(03), 393-412.

Abarbanel, H. D., Rozdeba, P. J., & Shirman, S. (2018). Machine Learning: Deepest Learning as Statistical Data Assimilation Problems. *Neural Computation*, *30*, 2025-2055.

Aleta, A., Martín-Corral, D., Piontti, A. P. y., Ajelli, M., Litvinova, M., Chinazzi, M., Dean, N. E., Halloran, M. E., Longini, I. M., Merler, S., Pentland, A., Vespignani, A., Moro, E., & Moreno, Y. (2020). Modeling the impact of social distancing, testing, contact tracing and household quarantine on second-wave scenarios of the COVID-19 epidemic. *medRxiv*.

Anderson, J. L. (2001). An ensemble adjustment Kalman filter for data assimilation. *Monthly weather review*, *129*(12), 2884-2903.

Anderson, J. L., & Anderson, S. L. (1999). A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts. *Monthly Weather Review*, *127*(12), 2741-2758.

Andrieu, C., & Doucet, A. (2003). Online expectation-maximization type algorithms for parameter estimation in general state space models. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03).*, *6*, VI-69.

Annan, J., & Hargreaves, J. (2004). Efficient parameter estimation for a highly chaotic system. *Tellus A: Dynamic Meteorology and Oceanography*, *56*(5), 520-526.

Arenas, A., Cota, W., Gómez-Gardeñes, J., Gómez, S., Granell, C., Matamalas, J. T., Soriano-Paños, D., & Steinegger, B. (2020). Modeling the Spatiotemporal Epidemic Spreading of COVID-19 and the Impact of Mobility and Social Distancing Interventions. *Phys. Rev. X*, *10*, 041055.

Arulampalam, M. S., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on signal processing*, *50*(2), 174-188.

Atkins, E., Morzfeld, M., & Chorin, A. J. (2013). Implicit particle methods and their connection with variational data assimilation. *Monthly Weather Review*, *141*(6), 1786-1803.

Berry, T., & Sauer, T. (2013). Adaptive ensemble Kalman filtering of non-linear systems. *Tellus A: Dynamic Meteorology and Oceanography*, *65*(1), 20331.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bishop, C. H. (2016). The GIGG-EnKF: ensemble Kalman filtering for highly skewed non-negative uncertainty distributions. *Quarterly Journal of the Royal Meteorological Society*, *142*(696), 1395-1412.

Bishop, C. H., Etherton, B. J., & Majumdar, S. J. (2001). Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Monthly weather review*, *129*(3), 420-436.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the national academy of sciences*, *99*(suppl 3), 7280-7287.

Burgers, G., Jan van Leeuwen, P., & Evensen, G. (1998). Analysis scheme in the ensemble Kalman filter. *Monthly weather review*, *126*(6), 1719-1724.

Cappé, O. (2009). Online sequential monte carlo EM algorithm. *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, 37-40.

Cappé, O. (2011). Online EM algorithm for hidden Markov models. *Journal of Computational and Graphical Statistics*, *20*(3), 728-749.

Carrassi, A., Bocquet, M., Bertino, L., & Evensen, G. (2018). Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *WIREs Climate Change*, *9*(5), e535.

Carrassi, A., Bocquet, M., Hannart, A., & Ghil, M. (2017). Estimating model evidence using data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *143*(703), 866-880.

Chorin, A. J., & Tu, X. (2009). Implicit sampling for particle filters. *Proceedings of the National Academy of Sciences*, *106*(41), 17249-17254.

Cocucci, T. J., Pulido, M., Lucini, M., & Tandeo, P. (2021). Model error covariance estimation in particle and ensemble Kalman filters using an online expectation–maximization algorithm. *Quarterly Journal of the Royal Meteorological Society*, *147*(734), 526-543.

Cocucci, T. J., Pulido, M., Aparicio, J. P., Ruiz, J., Simoy, M. I., & Rosa, S. (2022). Inference in epidemiological agent-based models using ensemble-based data assimilation. *PLOS ONE*, *17*(3), 1-28.

Cosme, E., Verron, J., Brasseur, P., Blum, J., & Auroux, D. (2012). Smoothing problems in a Bayesian framework and their linear Gaussian solutions. *Monthly Weather Review*, *140*(2), 683-695.

Courtier, P., Andersson, E., Heckley, W., Vasiljevic, D., Hamrud, M., Hollingsworth, A., Rabier, F., Fisher, M., & Pailleux, J. (1998). The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. *Quarterly Journal of the Royal Meteorological Society*, *124*(550), 1783-1807.

Daum, F., & Huang, J. (2009). Nonlinear filters with particle flow induced by log-homotopy. *Signal Processing, Sensor Fusion, and Target Recognition XVIII*, *7336*, 733603.

Dee, D. P. (1995). On-line estimation of error covariance parameters for atmospheric data assimilation. *Monthly weather review*, *123*(4), 1128-1145.

Dee, D. P., & Da Silva, A. M. (1999). Maximum-likelihood estimation of forecast and observation error covariance parameters. Part I: Methodology. *Monthly Weather Review*, *127*(8), 1822-1834.

DelSole, T., & Yang, X. (2010). State and parameter estimation in stochastic dynamical models. *Physica D: Nonlinear Phenomena*, *239*(18), 1781-1788.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1-22.

Desroziers, G., Berre, L., Chapnik, B., & Poli, P. (2005). Diagnosis of observation, background and analysis-error statistics in observation space. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, *131*(613), 3385-3396.

Doucet, A., De Freitas, N., Gordon, N. J., et al. (2001). *Sequential Monte Carlo methods in practice* (Vol. 1). Springer.

Dreano, D., Tandeo, P., Pulido, M., Ait-El-Fquih, B., Chonavel, T., & Hoteit, I. (2017). Estimating model-error covariances in nonlinear state-space models using Kalman smoothing and the expectation–maximization algorithm. *Quarterly Journal of the Royal Meteorological Society*, *143*(705), 1877-1885.

Epstein, J. M., & Axtell, R. (1996). *Growing artificial societies: social science from the bottom up*. Brookings Institution Press.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, *99*(C5), 10143-10162.

Evensen, G., Amezcua, J., Bocquet, M., Carrassi, A., Farchi, A., Fowler, A., Houtekamer, P. L., Jones, C. K., de Moraes, R. J., Pulido, M., Sampson, C., & Vossepoel, F. C. (2020). An international assessment of the COVID-19 pandemic using ensemble data assimilation. *medRxiv*.

Faes, C., Abrams, S., Van Beckhoven, D., Meyfroidt, G., Vlieghe, E., Hens, N., et al. (2020). Time between symptom onset, hospitalisation and recovery or death: statistical analysis of Belgian COVID-19 patients. *International journal of environmental research and public health*, *17*(20), 7560.

Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J. W., et al. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, *584*(7820), 257-261.

Frei, M., & Künsch, H. R. (2013). Bridging the ensemble Kalman and particle filters. *Biometrika*, *100*(4), 781-800.

Ghostine, R., Gharamti, M., Hassrouny, S., & Hoteit, I. (2021). An extended SEIR model with vaccination for forecasting the COVID-19 pandemic in Saudi Arabia using an ensemble Kalman filter. *Mathematics*, *9*(6), 636.

Golub, G. H., & Van Loan, C. F. (1996). Matrix Computations (3ª ed.).

Gordon, N. J., Salmond, D. J., & Smith, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE proceedings F (radar and signal processing)*, *140*(2), 107-113.

Grewal, M. S., & Andrews, A. P. (2010). Applications of Kalman Filtering in Aerospace 1960 to the Present [Historical Perspectives]. *IEEE Control Systems Magazine*, *30*(3), 69-78.

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., Thulke, H.-H., Weiner, J., Wiegand, T., & DeAngelis, D. L. (2005). Pattern-oriented modeling of agent-based complex systems: lessons from ecology. *science*, *310*(5750), 987-991.

Guan, J., Wei, Y., Zhao, Y., & Chen, F. (2020). Modeling the transmission dynamics of COVID-19 epidemic: a systematic review. *Journal of Biomedical Research*, *34*(6), 422.

Gupta, N. (2007). Kalman filtering in the presence of state space equality constraints. *2007 Chinese Control Conference*, 107-113.

Hamill, T. M., & Snyder, C. (2000). A hybrid ensemble Kalman filter–3D variational analysis scheme. *Monthly Weather Review*, *128*(8), 2905-2919.

Hamill, T. M., & Whitaker, J. S. (2005). Accounting for the error due to unresolved scales in ensemble data assimilation: A comparison of different approaches. *Monthly weather review*, *133*(11), 3132-3147.

Hamill, T. M., Whitaker, J. S., & Snyder, C. (2001). Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Monthly Weather Review*, *129*(11), 2776-2790.

Helbing, D. (2012). *Social self-organization: Agent-based simulations and experiments to study emergent social behavior*. Springer.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM review*, *42*(4), 599-653.

Hooten, M., Wikle, C., & Schwob, M. (2020). Statistical Implementations of Agent-Based Demographic Models. *International Statistical Review*, *88*(2), 441-461.

Ionides, E. L., Bretó, C., & King, A. A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, *103*(49), 18438-18443.

Ivorra, B., Ferrández, M. R., Vela-Pérez, M., & Ramos, A. M. (2020). Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. The case of China. *Communications in nonlinear science and numerical simulation*, *88*, 105303.

Jazwinski, A. H. (1970). *Stochastic processes and filtering theory*. Academic Press.

Jordan, M. I. (1999). *Learning in graphical models*. MIT press.

Kalman, R. E., & Bucy, R. S. (1961). New results in linear filtering and prediction theory.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems.

Kalnay, E. (2003). *Atmospheric modeling, data assimilation and predictability*. Cambridge university press.

Kalnay, E., Li, H., Miyoshi, T., Yang, S.-C., & Ballabrera-Poy, J. (2007). 4-D-Var or ensemble Kalman filter? *Tellus A: Dynamic Meteorology and Oceanography*, *59*(5), 758-773.

Katzfuss, M., Stroud, J. R., & Wikle, C. K. (2016). Understanding the ensemble Kalman filter. *The American Statistician*, *70*(4), 350-357.

Kermack, W. O., McKendrick, A. G., & Walker, G. T. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *115*(772), 700-721.

Kerr, C. C., Stuart, R. M., Mistry, D., Abeysuriya, R. G., Hart, G., Rosenfeld, K., Selvaraj, P., Núñez, R. C., Hagedorn, B., George, L., Izzo, A., Palmer, A., Delport, D., Bennette, C., Wagner, B., Chang, S., Cohen, J. A., Panovska-Griffiths, J., Jastrzębski, M., … Klein, D. J. (2020). Covasim: an agent-based model of COVID-19 dynamics and interventions. *medRxiv*.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klepac, P., Kissler, S., & Gog, J. (2018). Contagion! the bbc four pandemic–the model behind the documentary. *Epidemics*, *24*, 49-59.

Kovachki, N. B., & Stuart, A. M. (2019). Ensemble Kalman inversion: a derivative-free technique for machine learning tasks. *Inverse Problems*, *35*(9), 095005.

Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

LeGland, F., & Mevel, L. (1997). Recursive estimation in hidden Markov models. *Proceedings of the 36th IEEE Conference on Decision and Control*, *4*, 3468-3473.

Li, H., Kalnay, E., & Miyoshi, T. (2009). Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, *135*(639), 523-533.

Li, R., Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W., & Shaman, J. (2020). Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*, *368*(6490), 489-493.

Liu, J. S., & Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American statistical association*, *93*(443), 1032-1044.

Liu, Q. (2017). Stein variational gradient descent as gradient flow. *Advances in neural information processing systems*, *30*.

Liu, Q., & Wang, D. (2016). Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, *29*.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of atmospheric sciences*, *20*(2), 130-141.

Lorenz, E. N. (1996). Predictability: A problem partly solved. *Proc. Seminar on predictability*, *1*(1).

Lucini, M. M., van Leeuwen, P. J., & Pulido, M. (2021). Model Error Estimation Using the Expectation Maximization Algorithm and a Particle Flow Filter. *SIAM/ASA Journal on Uncertainty Quantification*, *9*(2), 681-707.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Mandel, J., Bennethum, L. S., Beezley, J. D., Coen, J. L., Douglas, C. C., Kim, M., & Vodacek, A. (2008). A wildland fire model with data assimilation. *Mathematics and Computers in Simulation*, *79*(3), 584-606.

Mehra, R. (1970). On the identification of variances and adaptive Kalman filtering. *IEEE Transactions on automatic control*, *15*(2), 175-184.

Miyoshi, T. (2011). The Gaussian Approach to Adaptive Covariance Inflation and Its Implementation with the Local Ensemble Transform Kalman Filter. *Monthly Weather Review*, *139*(5), 1519-1535.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Murray, J. D. (2007). *Mathematical Biology I: An Introduction* (3rd). Springer.

Neal, R. M. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and computing*, *6*(4), 353-366.

Neal, R. M., & Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. En *Learning in graphical models* (pp. 355-368). Springer.

Noble, J. V. (1974). Geographic and temporal development of plagues. *Nature*, *250*(5469), 726-729.

Petersen, K. B., & Pedersen, M. S. (2012). The Matrix Cookbook. http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html

Pulido, M., Tandeo, P., Bocquet, M., Carrassi, A., & Lucini, M. (2018). Stochastic parameterization identification using ensemble Kalman filtering combined with maximum likelihood methods. *Tellus A: Dynamic Meteorology and Oceanography*, *70*(1), 1-17.

Pulido, M., & van Leeuwen, P. J. (2019). Sequential Monte Carlo with kernel embedded mappings: The mapping particle filter. *Journal of Computational Physics*, *396*, 400-415.

Rabier, F., & Liu, Z. (2003). Variational data assimilation: theory and overview. *Proc. ECMWF Seminar on Recent Developments in Data Assimilation for Atmosphere and Ocean, Reading, UK, September 8–12*, 29-43.

Reich, S. (2013). A nonparametric ensemble transform method for Bayesian inference. *SIAM Journal on Scientific Computing*, *35*(4), A2013-A2024.

Roche, B., Drake, J. M., & Rohani, P. (2011). An Agent-Based Model to study the epidemiological and evolutionary dynamics of Influenza viruses. *BMC bioinformatics*, *12*(1), 1-10.

Ruchi, S., Dubinkina, S., & Iglesias, M. (2019). Transform-based particle filtering for elliptic Bayesian inverse problems. *Inverse Problems*, *35*(11), 115005.

Ruiz, J. J., Pulido, M., & Miyoshi, T. (2013a). Estimating Model Parameters with Ensemble-Based Data Assimilation: A Review. *Journal of the Meteorological Society of Japan. Ser. II*, *91*(2), 79-99.

Ruiz, J. J., Pulido, M., & Miyoshi, T. (2013b). Estimating Model Parameters with Ensemble-Based Data Assimilation: Parameter Covariance Treatment. *Journal of the Meteorological Society of Japan. Ser. II*, *91*(4), 453-469.

Särkkä, S. (2013). *Bayesian filtering and smoothing*. Cambridge University Press.

Shaman, J., & Karspeck, A. (2012). Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences*, *109*(50), 20425-20430.

Shaman, J., Karspeck, A., Yang, W., Tamerius, J., & Lipsitch, M. (2013). Real-time influenza forecasts during the 2012–2013 season. *Nature Communications*, *4*(1), 2837.

Shumway, R. H., & Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of time series analysis*, *3*(4), 253-264.

Silva, P. C., Batista, P. V., Lima, H. S., Alves, M. A., Guimarães, F. G., & Silva, R. C. (2020). COVID-ABS: An agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions. *Chaos, Solitons & Fractals*, *139*, 110088.

Simoy, M. I., & Aparicio, J. P. (2021). Socially structured model for COVID-19 pandemic: design and evaluation of control measures. *In press*.

Stordal, A. S., Karlsen, H. A., Nævdal, G., Skaug, H. J., & Vallès, B. (2011). Bridging the ensemble Kalman filter and particle filters: the adaptive Gaussian mixture filter. *Computational Geosciences*, *15*(2), 293-305.

Stroud, J. R., Katzfuss, M., & Wikle, C. K. (2018). A Bayesian adaptive ensemble Kalman filter for sequential state and parameter estimation. *Monthly weather review*, *146*(1), 373-386.

Talagrand, O., & Courtier, P. (1987). Variational assimilation of meteorological observations with the adjoint vorticity equation. I: Theory. *Quarterly Journal of the Royal Meteorological Society*, *113*(478), 1311-1328.

Tandeo, P., Ailliot, P., Bocquet, M., Carrassi, A., Miyoshi, T., Pulido, M., & Zhen, Y. (2020). A review of innovation-based methods to jointly estimate model and observation error covariance matrices in ensemble data assimilation. *Monthly Weather Review*, *148*(10), 3973-3994.

Tandeo, P., Pulido, M., & Lott, F. (2015). Offline parameter estimation using EnKF and maximum likelihood error covariance estimates: Application to a subgrid-scale orography parametrization. *Quarterly journal of the royal meteorological society*, *141*(687), 383-395.

Tesfatsion, L., & Judd, K. L. (2006). *Handbook of computational economics: agent-based computational economics*. Elsevier.

Van Leeuwen, P. J., Künsch, H. R., Nerger, L., Potthast, R., & Reich, S. (2019). Particle filters for high-dimensional geoscience applications: A review. *Quarterly Journal of the Royal Meteorological Society*, *145*(723), 2335-2365.

Verity, R., Okell, L. C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., Cuomo-Dannenburg, G., Thompson, H., Walker, P., Fu, H., et al. (2020). Estimates of the severity of COVID-19 disease. *MedRxiv*.

Vossepoel, F. C., & Jan van Leeuwen, P. (2007). Parameter estimation using a particle method: Inferring mixing coefficients from sea level observations. *Monthly weather review*, *135*(3), 1006-1020.

Vynnycky, E., & White, R. (2010). *An introduction to infectious disease modelling*. OUP oxford.

Wainwright, M. J., & Jordan, M. I. (2008). *Graphical models, exponential families, and variational inference*. Now Publishers Inc.

Ward, J. A., Evans, A. J., & Malleson, N. S. (2016). Dynamic calibration of agent-based models using data assimilation. *Royal Society Open Science*, *3*(4), 150703.

Wasserman, L. (2004). *All of statistics: a concise course in statistical inference* (Vol. 26). Springer.

Whitaker, J. S., & Hamill, T. M. (2002). Ensemble data assimilation without perturbed observations. *Monthly weather review*, *130*(7), 1913-1924.

Wu, C. J. (1983). On the convergence properties of the EM algorithm. *The Annals of statistics*, 95-103.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zhao, H., & Feng, Z. (2020). Staggered release policies for COVID-19 control: Costs and benefits of relaxing restrictions by age and risk. *Mathematical biosciences*, *326*, 108405.

Zhu, M., Van Leeuwen, P. J., & Amezcua, J. (2016). Implicit equal-weights particle filter. *Quarterly Journal of the Royal Meteorological Society*, *142*(698), 1904-1919.