

Curso D

**Esteban Tabak**

New York University, EEUU

### Análisis de datos a través del transporte óptimo

Este curso describe una metodología general para determinar, en base a observaciones, cómo un grupo de variables depende de otras. Una forma de representar esta dependencia es a través de una probabilidad condicional  $p(x|z)$ , donde  $x$  comprende las variables de interés y  $z$  los factores que las afectan, que pueden ser conocidos o cuya determinación puede ser parte del problema.

Un ejemplo importante de aplicación consiste en determinar el efecto que un tratamiento bajo consideración pueda tener en un paciente específico. Aquí  $x$  representa el estado de salud después del tratamiento (presión arterial, número de glóbulos rojos, nivel de azúcar en sangre) y los factores  $z$  incluyen características medibles del paciente (edad, peso, estado de salud actual), características del tratamiento considerado (tipo, dosis) y características no conocidas, “idiosincráticas”, del paciente, que determinan su reacción individual a un tratamiento. Otro ejemplo es el pronóstico del tiempo, que estima  $x$ , el estado futuro del tiempo, en función de  $z$ : lugar, día y hora, el estado del tiempo actual, el índice del Niño. Determinar la existencia de factores ocultos, tal como precisamente El Niño, ilustra otra faceta del problema.

El problema matemático de transporte óptimo busca el mapa  $y = T(x)$  que transforma una distribución  $p(x)$  en otra  $m(y)$  minimizando un “costo de transporte”, que para nosotros representa una medida de la deformación de los datos. En este curso, veremos como este problema, extendido a transportes condicionales  $y = T(x | z)$ , presenta una perspectiva natural y una metodología efectiva para determinar probabilidades condicionales y simularlas (extraer muestras), así como descubrir factores ocultos. Estudiaremos tanto la parte conceptual del problema como su formulación matemática e implementación en términos de observaciones. Veremos también cómo esta formulación generaliza herramientas estadísticas clásicas, tales como regresión, separación en grupos y componentes principales.

El curso no requiere prerrequisitos particulares, más que conocimiento de cálculo, álgebra lineal y probabilidad básica.