Algunas aplicaciones y perspectivas de la optimización Robusta Distribucional (DRO) con métrica de Wasserstein

Diego Fonseca

30 de junio de 2021

Preliminares

Maquinas de Soporte Vectorial SVM

Optimización de portafolios con Conditional Value at Risk (CVaR)

Optimización de portafolios desde la perspectiva de Markowitz

Determinación del parámetro bandwidth en le estimación de funciones de densidad por kernels

Preliminares

Métrica de Wasserstein

Definición 1 (Métrica de Wassertein)

La *distancia de Wasserstein* $W_p(\mu, \nu)$ entre $\mu, \nu \in \mathcal{P}_p(\Xi)$[1] es definida por

$$W_p^p(\mu,\nu) := \inf_{\Pi \in \mathcal{P}(\Xi \times \Xi)} \left\{ \int_{\Xi \times \Xi} d^p(\xi,\zeta)\Pi(d\xi, d\zeta) \;\middle|\; \begin{array}{l} \Pi(\cdot \times \Xi) = \mu(\cdot), \\ \Pi(\Xi \times \cdot) = \nu(\cdot) \end{array} \right\}$$

donde

$$\mathcal{P}_p(\Xi) := \left\{ \mu \in \mathcal{P}(\Xi) \;:\; \int_{\Xi} d^p(\xi,\zeta_0)\mu(d\xi) < \infty \text{ para algun } \zeta_0 \in \Xi \right\}$$

donde d es una métrica en Ξ.

[1] La métrica p-Wasserstein también esta definida para distribuciones fuera de $\mathcal{P}_p(\Xi)$, lo que probablemente podría ocurrir es que ese conjunto la métrica de Wasserstein sea infinito.

Un objeto matemático importante en el contexto del presente trabajo son las bolas respecto a alguna métrica p-Wasserstein, en el contexto de $\mathcal{P}_p(\Xi)$ la bola de radio $\varepsilon > 0$ con centro en $\mu \in \mathcal{P}_p(\Xi)$ es

$$\mathcal{B}_\varepsilon^p(\mu) = \left\{ \nu \in \mathcal{P}_p(\Xi) \;\middle|\; W_p^p(\mu, \nu) \leq \varepsilon^p \right\}. \tag{1}$$

Un problema de optimización estocástica es de la forma

$$J^* = \min_{x \in \mathbb{X}} \mathbb{E}_{\mathbb{P}}[f(x, \xi)]$$

donde $f : \mathbb{X} \times \Xi \to \mathbb{R}$, $\mathbb{X}$ es la región factible y ξ es un elemento aleatorio con distribución $\mathbb{P}$ soportada en Ξ.

Aproximación robusta distribucional con métrica de Wasserstein de J^*

Sea $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ una muestra de $\mathbb{P}$ y $\widehat{\mathbb{P}}_N$ la distribución empírica determinada por esta muestra. Dado $\varepsilon > 0$ sea $\mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_N\right)$ la bola cerrada respecto a una métrica p-Wasserstein de radio ε y centro $\widehat{\mathbb{P}}_N$. El DRO que aproxima J^* superiormente con alta probabilidad es

$$\widehat{J}_N := \min_{x \in \mathbb{X}} \sup_{\mathbb{Q} \in \mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_N\right)} \mathbb{E}_{\mathbb{Q}}[f(x, \xi)]. \tag{2}$$

Suposición 1: Asumimos f que alguna de las siguientes condiciones:

1. f es continua y es tal que existe $C > 0$ y $\xi_0 \in \Xi$ tal que $|f(\xi)| \leq C(1 + d^p(\xi, \xi_o))$ para todo $\xi \in \Xi$.
2. f es acotada.
3. f es máximo de funciones concavas, es decir, $f = \text{máx}_{k \leq K} f_k$ donde cada $-f_k$ es propia, convexa e inferiormente semicontinua respecto a ξ.

Teorema 2

Bajo la Suposición 1 el problema (2) se puede reformular como el problema de optimización semi-infinito

$$\begin{cases} \inf\limits_{\lambda,\, x,\, s} & \lambda\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \textit{sujeto a} & \sup\limits_{\xi\in\Xi}\Big(f(x,\xi) - \lambda d^p(\langle x,\xi\rangle, \langle x,\widehat{\xi}_i\rangle)\Big) \le s_i \quad \forall i \le N. \\ & x \in \mathbb{X}, \\ & \lambda \ge 0. \end{cases} \tag{3}$$

Corolario 3

Asumimos $\Xi = \{\xi \in \mathbb{R}^m \ : \ C\xi \leq g\}$ *donde* C *es una matriz y* g *un vector de dimensiones apropiadas,* $p = 1$ *y* d *como una norma* $\|\cdot\|$ *en* $\mathbb{R}^m$ *y se consideran las funciones* $a_k(x, \xi) := \langle a_k x, \xi \rangle + b_k$ *para todo* $k \leq K$*. Si* $f(x, \xi) = \max_{k \leq K} a_k(x, \xi)$ *entonces el valor óptimo de (2) es igual al valor óptimo de*

$$\begin{cases} \inf\limits_{\lambda,\, x,\, s, \gamma_{ik}} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i & \\ \text{sujeto a} & b_k + \langle a_k x, \widehat{\xi}_i \rangle + \langle \gamma_{ik}, d - C\widehat{\xi}_i \rangle \leq s_i & \forall i \leq N \, \forall k \leq K. \\ & \left\| C^T \gamma_{ik} - a_k x \right\|_* \leq \gamma & \forall i \leq N \, \forall k \leq K. \\ & \gamma_{ik} \geq 0. & \end{cases} \tag{4}$$

Proposición 4

Sean ε *y* $\beta \in (0,1)$ *tales que* $\mathbb{P}^N \left\{ W_p(\mathbb{P}, \widehat{P}_N) \leq \varepsilon \right\} \geq 1 - \beta$. *Si* $\widehat{x}_N$ *una solución óptima de (2) entonces*

$\mathbb{P}^N \left\{ \mathbb{E}_P \left[f(\widehat{x}_N, \xi) \right] \leq \widehat{J}_N \right\} \geq 1 - \beta.$

Noten que siempre se tiene $J^* \leq \mathbb{E}_P \left[f(\widehat{x}_N, \xi) \right]$.

Maquinas de soporte vectorial

Maquinas de Soporte Vectorial SVM es una técnica de clasificación que consiste en encontrar un hiperplano en el espacio de características que separa los datos con amplio margen.
Consideremos el problema de clasificación binario, donde un numero finito de datos de entrenamiento $\{(\mathbf{x}_j, y_j)\}_{j=1}^{N}$ $subset\mathbb{R}^n \times \{-1, 1\}$ son dados y necesitamos encontrar un clasificador lineal, $f(\mathbf{x}, \mathbf{w}, b) = \text{sign}\left(\mathbf{w}^T\mathbf{x} + b\right)$. Supongamos que los datos no son linealmente separables.
SVM consiste en encontrar un vector w y un escalar b solucionando el problema cuadrático

$$\begin{cases} \min\limits_{\mathbf{w},b,\mathbf{s}} & \dfrac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum\limits_{j=1}^{N} s_j & \\ \text{sujeto a} & 1 - y_j(\mathbf{w}^T\mathbf{x}_j + b) \leq s_j & \forall j = 1, \ldots, N. \\ & s_j \geq 0 & \forall j = 1, \ldots, N. \end{cases} \tag{5}$$

Si definimos $h(\mathbf{x}, y; w, b) := \text{máx}\left\{1 - y_j(\mathbf{w}^T\mathbf{x} + b), 0\right\}$ y consideramos $\widehat{C} := NC$, entonces (5) se puede reformular como:

$$\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{\widehat{C}}{N}\sum_{j=1}^{N} h(\mathbf{x}_j, y_j; w, b). \tag{6}$$

Sea $\xi := (\mathbf{x}_j, y_j)$ representando el para de el vector aleatorio de características y el nivel de clasificación con distribución $\mathbb{P}$ y soporte es $\Xi := \mathcal{X} \times \{-1, 1\}$. Considere el problema de optimización convexo

$$\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \widehat{C} \int_{\Xi} h(\xi; w, b)\mathbb{P}(d\xi). \tag{7}$$

Note que (6) es la versión muestral (Sample Average Approximation SAA) de (7). Entonces, el problema central en realidad es (7), pero debido al desconocimiento de $\mathbb{P}$ en [2] se propone una aproximación robusta del problema.

Suponga que $\Xi := \mathcal{X} \times \{-1, 1\}$ esta equipado con una métrica d, asumimos que esa métrica satisface que para cualquier $\xi^1 = (\mathbf{x}^1, y^1), \xi^2 = (\mathbf{x}^2, y^2) \in \Xi$ se tiene

$$d(\xi^1, \xi^2) = d_{\overline{x}}(\mathbf{x}^1, \mathbf{x}^2) + d_{\overline{y}}(y^1, y^2)$$

donde d_y es definida como

$$d_{\overline{y}}(y^1, y^2) = \begin{cases} \delta & \text{si } y^1 \neq y^2 \\ 0 & \text{si } y^1 = y^2 \end{cases}$$

El parámetro δ determina la importancia de la distancia entre dos clases comparado a la distancia en el espacio de características. Consideramos la métrica 1-Wasserstein con función de costo d y denotamos por $\mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_N\right)$ la bola cerrada respecto a la métrica 1-Wasserstein con centro en la medida empírica y radio $\varepsilon > 0$.

La versión robusta distribucional de (7) que notaremos DR-SVM es

$$\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \widehat{C} \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_N\right)} \int_{\Xi} h(\xi; w, b)\mathbb{Q}(d\xi). \tag{8}$$

Abordar el problema anterior es motivado por situaciones donde un numero pequeño de muestras están disponibles ó cuando los datos coleccionados están afectados por ruido, en tales casos la versión muestral no representa adecuadamente la distribución $\mathbb{P}$.

Por el Teorema 2 el problema interno del DR-SVM

$$\sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon\left(\widehat{\mathbb{P}}_N\right)} \int_{\Xi} h(\xi; w, b)\mathbb{Q}(d\xi)$$

es equivalente a

$$\begin{cases} \inf\limits_{\lambda,\, x,\, s} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{(\mathbf{x},y)\in\Xi} \Big(1 - y_j(\mathbf{w}^T\mathbf{x} + b) - \lambda\left(d_{\overline{x}}(\mathbf{x}, \mathbf{x}_j) - d_{\overline{y}}(y, y_j)\right)\Big) \leq s_j \quad \forall j \leq N. \\ & x \in \mathbb{X}, \\ & \lambda \geq 0. \end{cases} \tag{9}$$

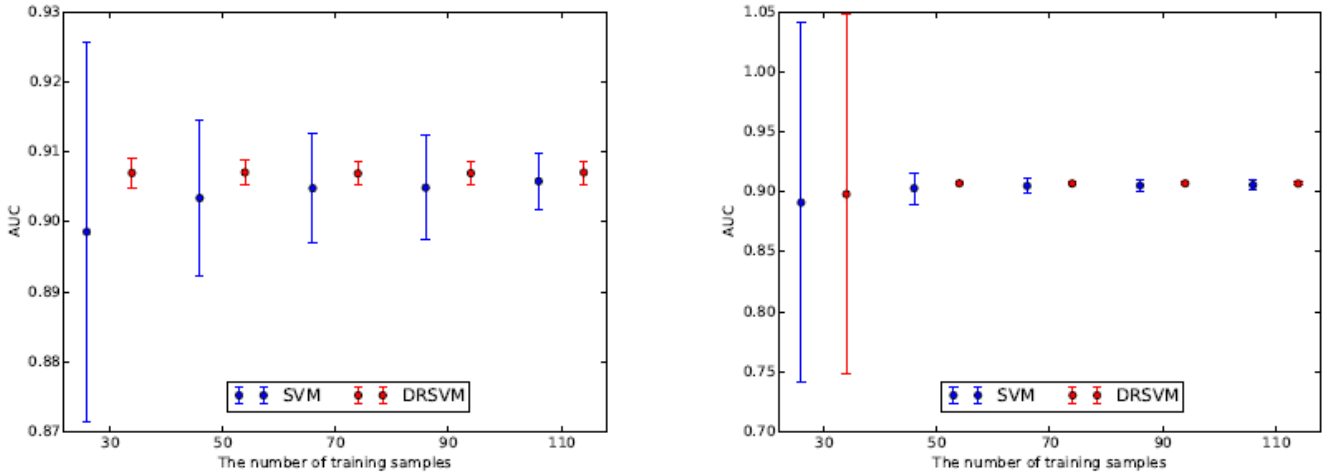
Este ultimo problema de optimización semi-infinita se soluciona con un algoritmo cutting-plane propuesto en [2].

Resultados para datos de prueba

Para la prueba inicial se generan aleatoriamente 50 puntos para el conjunto de entrenamiento y otros 1000 para los datos de prueba. Se entrenan los modelos SVM y DRO-SVM y se calcula la medida Area Bajo la Curva (AUC) de la curva ROC generada desde el modelo entrenado y los datos de prueba. Repetimos el experimento 100 veces para obtener un promedio y los errores estándar de la medida AUC para SVM y DR-SVM. Para este experimento se usa $\widehat{C} = 150$, $\varepsilon = 0{,}1$ y $\delta = 0{,}1$ y y se prueban los casos en donde $d_{\overline{x}}$ es la norma L-1 y L-∞.

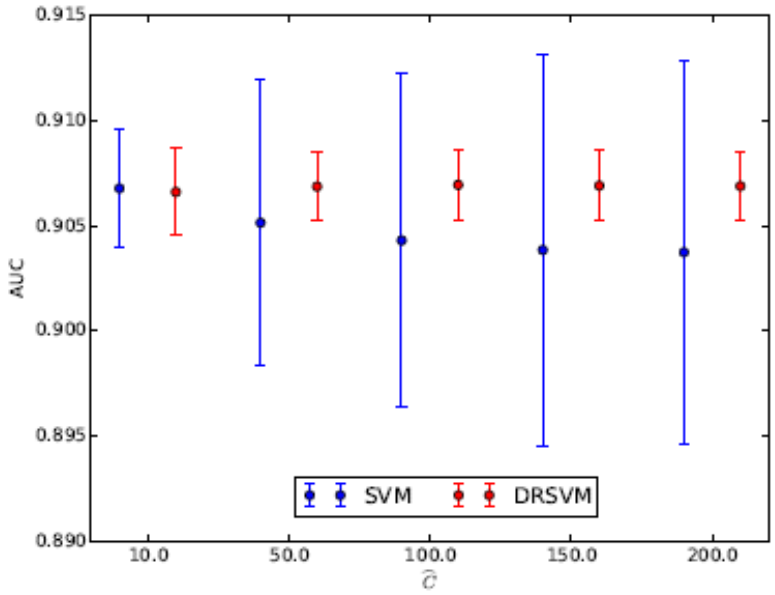
	AUC	(S.E.)
SVM	0.9039	0.0046
DR-SVM with L_1-norm	0.9069	0.0008
DR-SVM with L_∞-norm	0.9069	0.0008

Repetimos el experimento anterior para hallar intervalos de confianza para AUC pero en este caso variamos N.

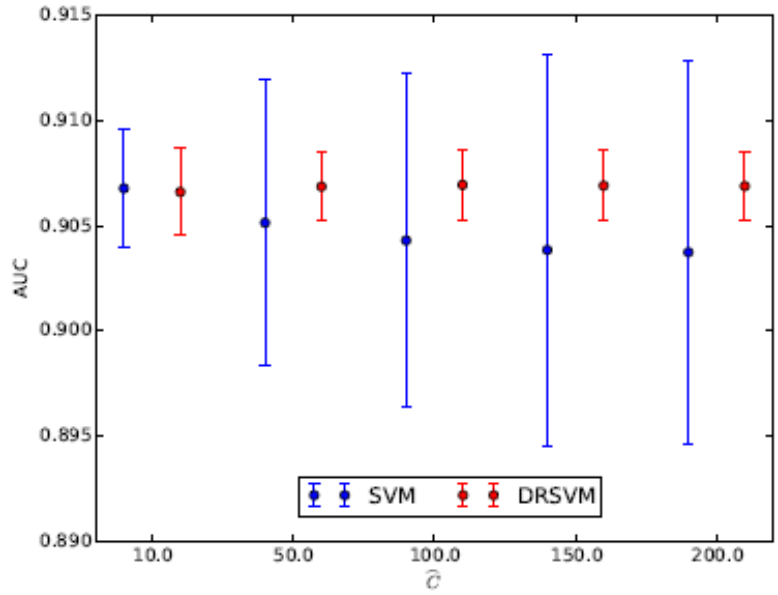


(a) L_1-norm (b) L_∞-norm

Figura: Análisis de sensibilidad para el numero de datos de entrenamiento N con intervalos de confianza del 95 %.

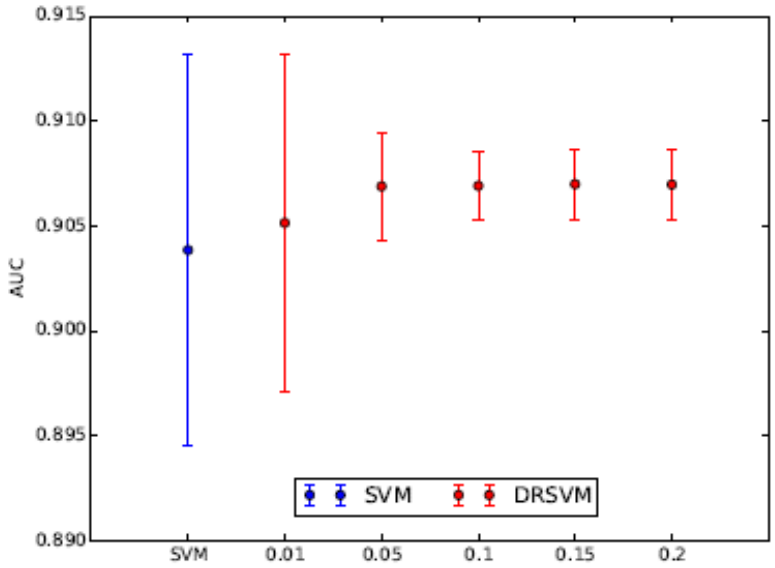


(a) L_1-norm

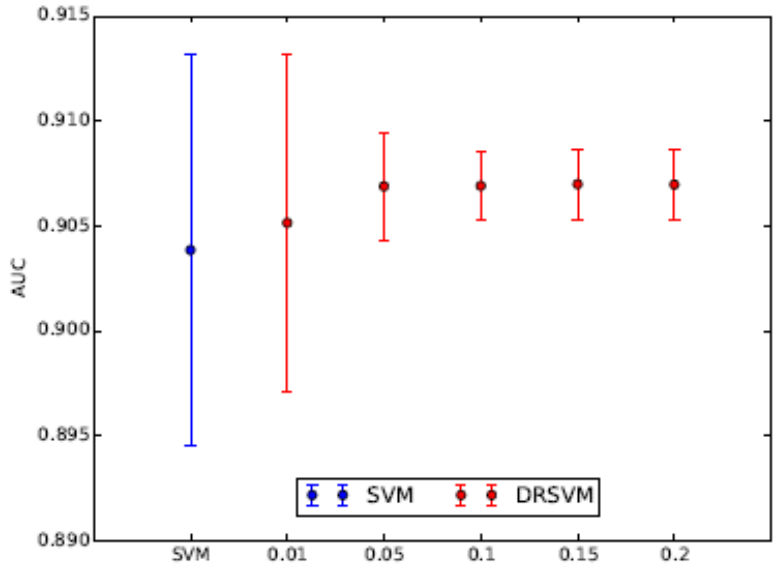


(b) L_∞-norm

Figura: Análisis de sensibilidad para $\widehat{C}$ ($\varepsilon = 0{,}1$, $\delta = 0{,}1$) con intervalos de confianza del 95 %.



(a) L_1-norm



(b) L_∞-norm

Figura: Análisis de sensibilidad para ε ($\widehat{C} = 150$, $\delta = 0{,}1$) con intervalos de confianza del 95 %.

Optimización de portafolios con Conditional Value at Risk (CVaR)

Considere un mercado de capitales consistiendo de m bienes cuyos retornos anuales son capturados por el vector aleatorio $\xi = [\xi_1, \ldots, \xi_m]^T$ cuya distribución es $\mathbb{P}$. Si las ventas cortas están prohibidas, un portafolio es codificado por un vector de pesos porcentuales $x = [x_1, \ldots, x_m]^T$ perteneciendo al conjunto

$$\mathbb{X} = \left\{ x \in \mathbb{R}^m_+ \; : \; \sum_{i=1}^{m} x_i = 1 \right\}.$$

El retorno del portafolio x es $\langle x, \xi \rangle$.

El objetivo es solucionar el problema de optimización estocástica:

$$J^* = \inf_{x \in \mathbb{X}} \left\{ \mathbb{E}_{\mathbb{P}} \left[-\langle x, \xi \rangle \right] + \rho \mathbb{P} CVaR_{\alpha} \left(-\langle x, \xi \rangle \right) \right\} \tag{10}$$

donde

$$\mathbb{P} CVaR_{\alpha} \left(-\langle x, \xi \rangle \right) := \inf_{\tau \in \mathbb{R}} \mathbb{E}_{\mathbb{P}} \left[\tau + \frac{1}{\alpha} \max \left\{ -\langle x, \xi \rangle - \tau, 0 \right\} \right].$$

En el problema (10) se minimiza la suma ponderada del valor esperado y el conditional value-at-risk (CVaR) de las perdidas del portafolio $-\langle x, \xi \rangle$, donde $\alpha \in (0, 1]$ es referida como el nivel de confianza del CVaR, y $\rho \in \mathbb{R}_+$ cuantifica la aversión al riesgo del inversionista.

Reemplazando en (10) el CVaR por su definición formal se obtiene que (10) se puede expresar como

$$J^* = \left\{ \mathbb{E}_{\mathbb{P}}[-\langle x, \xi \rangle] + \rho \inf_{\tau \in \mathbb{R}} \mathbb{E}_{\mathbb{P}} \left[\tau + \frac{1}{\alpha} \max \{ -\langle x, \xi \rangle - \tau, 0 \} \right] \right\}$$

$$= \inf_{x \in \mathbb{X}, \tau \in \mathbb{R}} \mathbb{E}_{\mathbb{P}} \left[\max_{k \leq K} (a_k \langle x, \xi \rangle + b_k \tau) \right]$$

donde $K = 2$, $a_1 = -1$, $a_2 = -1 - \frac{\rho}{\alpha}$, $b_1 = \rho$ y $b_2 = \rho \left(1 - \frac{1}{\alpha}\right)$.
El inversionista desconoce $\mathbb{P}$ pero tiene una muestra $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ de $\mathbb{P}$ y sabe que el soporte de $\mathbb{P}$ esta contenido en $\Xi = \{\xi \in \mathbb{R}^m \ : \ C\xi \leq d\}$.

La versión robusta distribución de (10) es

$$\widehat{J}_N(\varepsilon) := \inf_{x\in\mathbb{X},\, \tau\in\mathbb{R}} \sup_{\mathbb{Q}\in\mathcal{B}_\varepsilon(\widehat{P}_N)} \mathbb{E}_{\mathbb{Q}} \left[\max_{k\leq K}(a_k\langle x,\xi\rangle + b_k\tau)\right]. \qquad (11)$$

Por el Corolario 3 se tiene

$$\widehat{J}_N(\varepsilon) := \begin{cases} \inf\limits_{\lambda,\, x,\, s, \gamma_{ik}} & \lambda\varepsilon + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & b_k\tau + a_k\langle x, \widehat{\xi}_i\rangle + \langle \gamma_{ik}, d - C\widehat{\xi}_i\rangle \leq s_i \quad \forall i \leq N\, \forall k \leq K. \\ & \left\| C^T\gamma_{ik} - a_k x\right\|_* \leq \gamma \qquad\qquad\qquad\quad \forall i \leq N\, \forall k \leq K. \\ & \gamma_{ik} \geq 0. \end{cases} \qquad (12)$$

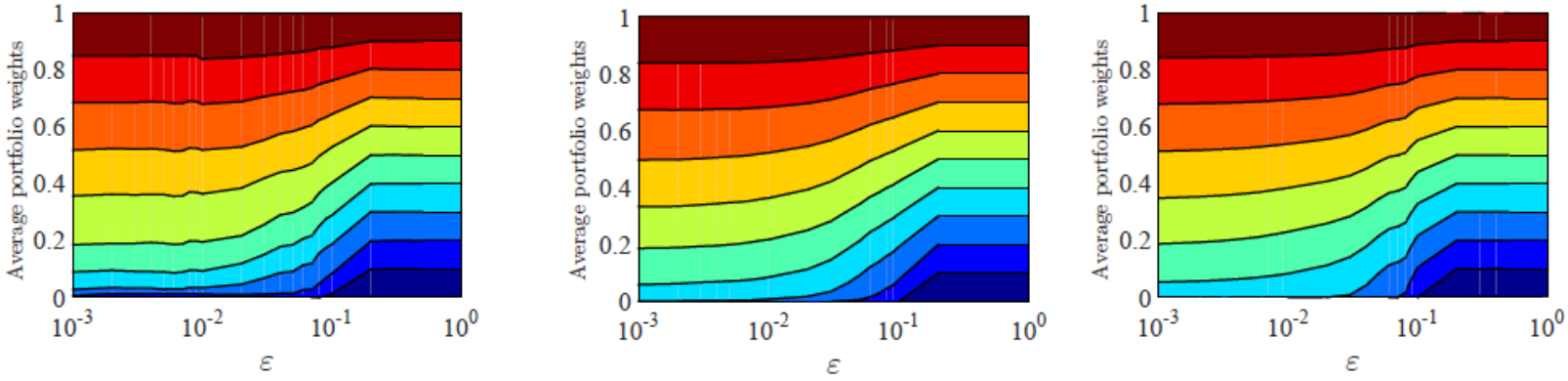
Proposición 5

Si $\{\varepsilon_k\}_{k\in\mathbb{N}} \subset \mathbb{R}_+$ *tiende a infinito y* $\{\widehat{x}_N(\varepsilon_k)\}_{k\in\mathbb{N}}$ *es la respectiva sucesión de soluciones óptimas de (11). Entonces* $\{\widehat{x}_N(\varepsilon_k)\}_{k\in\mathbb{N}}$ *converge al portafolio* $x^* = \frac{1}{m}e$ *siempre y cuando* Ξ *sea dado por:*

(i) $\Xi = \mathbb{R}^m$.

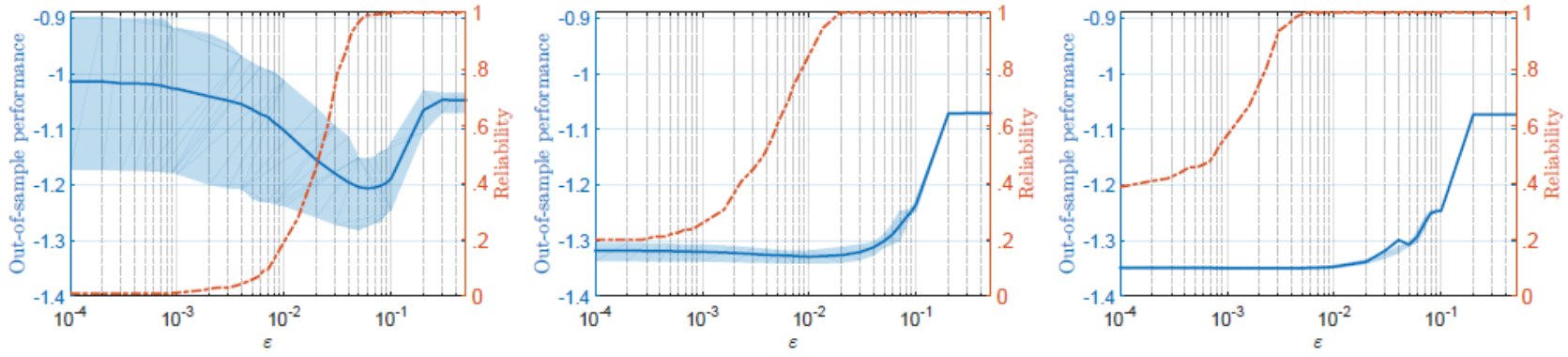
(ii) $\Xi = \{\xi \in \mathbb{R}^m \,:\, \xi \geq -e\}$.

donde e es el vector de unos.



(a) (b) (c)

Figura: Composición óptima del portafolio (compuesto de 10 bienes) como función del radio ε promediado sobre 200 simulaciones; los portafolios son pintados en orden ascendente, es decir, el peso del bien 1 en la parte baja en azul oscuro y al peso del bien 10 en la parte superior en rojo oscuro. (a) $N = 30$, (b) $N = 300$ y (c) $N = 3000$



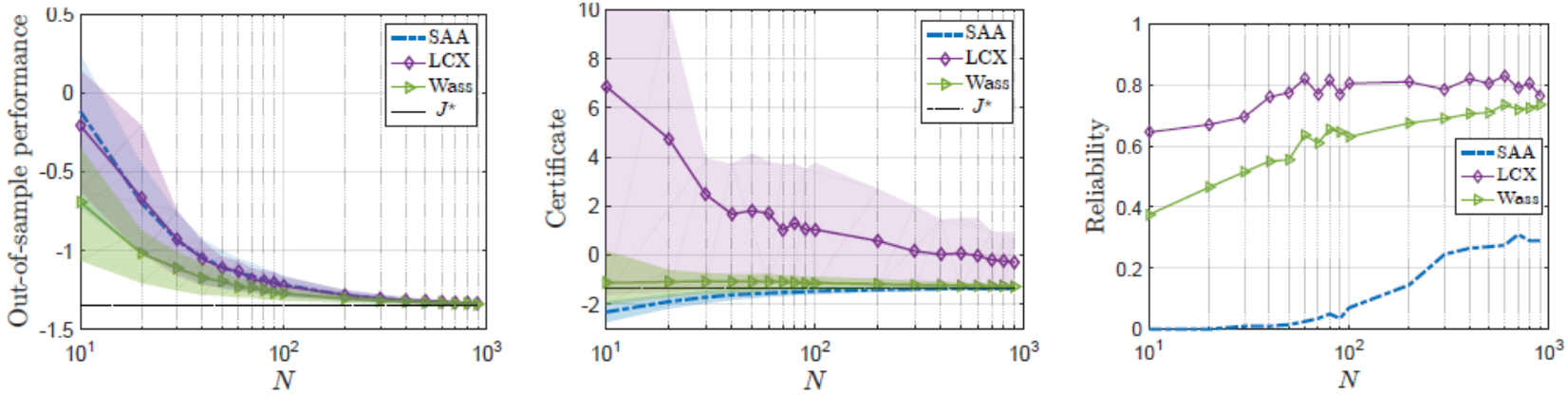
(a) $N = 30$ training samples (b) $N = 300$ training samples (c) $N = 3000$ training samples

Figura: Out-of-sample performance $J(\widehat{x}_N(\varepsilon))$ (eje izquierdo, linea solida y área sombreada) y reliability (confiabilidad) $\mathbb{P}^N\left[J(\widehat{x}_N(\varepsilon)) \leq \widehat{J}_N(\varepsilon)\right]$ (eje derecho, linea punteada) como función de el radio ε y estimado en 200 simulaciones.

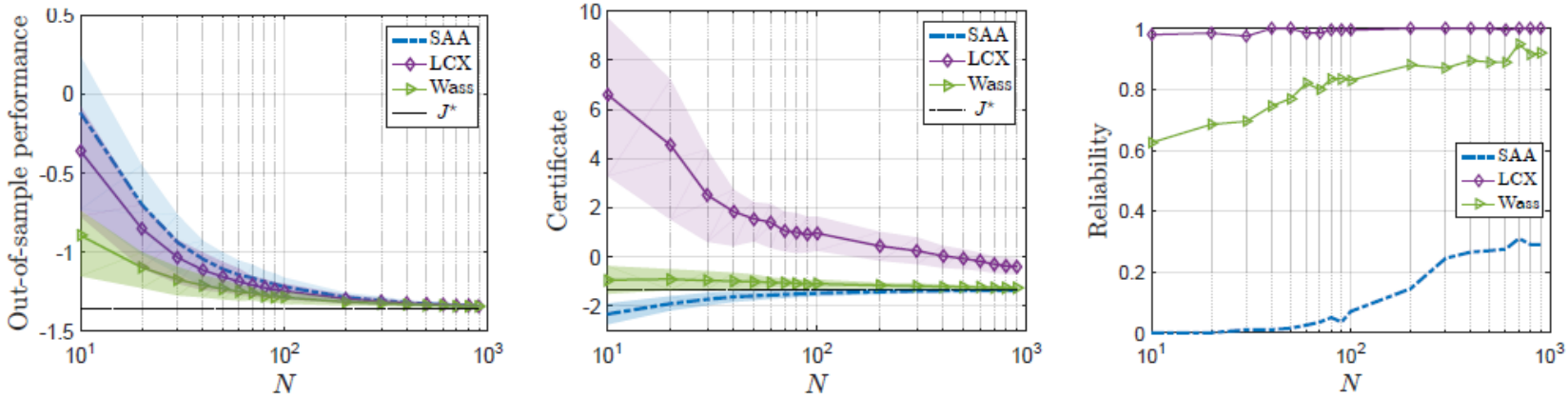
Elección de ε:

Diferentes radio ε pueden resultar en diferentes portafolios robustos $\widehat{x}_N(\varepsilon)$ con muy diferentes out-of-sample performance $J(\widehat{x}_N(\varepsilon))$. Idealmente, uno debería seleccionar el radio $\widehat{\varepsilon}_N^{\mathrm{opt}}$ que minimice $J(\widehat{x}_N(\varepsilon))$ sobre todo $\varepsilon \geq 0$. Pero J depende de $\mathbb{P}$ que es desconocida, en la practica, lo mejor es aproximar $\widehat{\varepsilon}_N^{\mathrm{opt}}$ a partir de los datos, dos métodos para lograr este objetivos son:

- **Holdout method:** Se divide la muestra $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ en un conjunto de entrenamiento de tamaño N_T y un conjunto de validación de tamaño $N_V = N - N_T$. Usando únicamente el conjunto de entrenamiento, se soluciona (11) para un conjnto garande pero finito de candidatos a ser ε y por cada candidato se obtiene $\widehat{x}_N(\varepsilon)$. Se usa el conjunto de validación para estimar $J(\widehat{x}_N(\varepsilon))$ muestralmente por cada candidato. El mejor candidato sera $\widehat{\varepsilon}_N^{\mathrm{hm}}$ que es cualquier de los candidatos ε que minimice la cantidad anterior.
- **k-fold cross validation:** Se divide la muestra $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ en k subconjuntos, y se ejecuta el holdout method K veces. En cada ejecución se usa exactamente uno de los subconjuntos como conjunto de validación y la unión de los $k-1$ restantes subconjuntos como conjunto de entrenamiento. El mejor candidato sera $\widehat{\varepsilon}_N^{\mathrm{xv}}$ que es el promedio de los ε obtenidos en las k ejecuciones del Holdout method.



(a) Holdout method (b) Holdout method (c) Holdout method



(d) k-fold cross validation (e) k-fold cross validation (f) k-fold cross validation

Figura: Out-of-sample performance $J(\widehat{x}_N(\varepsilon))$, certificado $\widehat{J}_N$ y reliability (confiabilidad) $\mathbb{P}^N\left[J(\widehat{x}_N(\varepsilon)) \leq \widehat{J}_N(\varepsilon)\right]$ para los soluciones SAA, LCX y Wasserstein como función de N.

Optimización de portafolios desde la perspectiva de Markowitz

Siguiendo con el contexto de la sección anterior los siguientes conceptos son los que tomaran relevancia en esta sección:

$$\text{Retorno esperado} := \mathbb{E}_{\mathbb{P}}\left[\langle x, \xi \rangle\right].$$

$$\text{Volatilidad} := \mathrm{Var}_{\mathbb{P}}\left[\langle x, \xi \rangle\right] = \mathbb{E}_{\mathbb{P}}\left[\left(\langle x, \xi \rangle - \mathbb{E}_{\mathbb{P}}\left[\langle x, \xi \rangle\right]\right)^2\right].$$

Lo ideal es encontrar un vector de pesos x que le garantice un retorno esperado alto pero con una volatilidad baja, en tal virtud el inversionista establece un nivel de retorno esperado mínimo μ, esto significa que los únicos vectores de pesos x que considerará son aquellos que garantizan un retorno esperado igual o mayor que μ. Esta visión es representada en el siguiente modelo:

$$J := \begin{cases} \min\limits_{x \in \mathbb{R}^m} & \mathrm{Var}_{\mathbb{P}}\left[\langle x, \xi \rangle\right] \\ \text{sujeto a} & \mathbb{E}_{\mathbb{P}}\left[\langle x, \xi \rangle\right] \geq \mu, \\ & \sum_{i=1}^{m} x_i = 1. \\ & x_i \geq 0. \end{cases} \tag{13}$$

Si se conociera la matriz de covarianza E y el vector de valores esperados $\mathbf{m}$ del vector aleatorio ξ entonces (13) es equivalente al problema de optimización

$$\begin{cases} \min\limits_{x \in \mathbb{R}^m} & x^T E x \\ \text{sujeto a} & \mathbf{m}^T x \geq \mu, \\ & \sum_{i=1}^{m} x_i = 1. \\ & x_i \geq 0. \end{cases} \tag{14}$$

Pero en la practica E y $\mathbf{m}$ no son conocidos, ante esta situación es común considerar E y $\mathbf{m}$ como las versiones muéstrales. De modo que la aproximación Robusta Distribucional es una opción viable.

Versión Robusta Distribucional:

Fijando $x \in \mathbb{X}$ definimos $\zeta^x := \langle x, \xi \rangle$ la cual es una variable aleatoria, llamamos $\mathbb{P}^x$ su distribución la cual depende de $\mathbb{P}$, luego, dada $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ una muestra de $\mathbb{P}$, entonces $\widehat{\zeta}_1^x, \ldots, \widehat{\zeta}_N^x$ definida por $\widehat{\zeta}_i^x := \langle x, \widehat{\xi}_i \rangle$ es una muestra de ζ^x, esto permite definir la distribución empírica $\widehat{\mathbb{P}}_N^x$ asociada a ζ^x la cual es dada por

$$\widehat{P}_N^x := \sum_{i=1}^{N} \delta_{\widehat{\zeta}_i^x}.$$

En el espació de las distribuciones se considera la noción de distancia determinada por la métrica de 2-Wasserstein denotada por W_2, esto permite considerar el conjunto

$$\mathcal{B}_{\varepsilon\|x|}(\widehat{\mathbb{P}}_N^x) = \left\{ \mathbb{Q} \in \mathcal{P}(\mathbb{R}^m) \,\middle|\, W(\mathbb{Q}, \widehat{\mathbb{P}}_N^x) \leq \varepsilon \|x\| \right\}.$$

Este conjunto es la bola respecto a la métrica 2-Wasserstein con centro en $\widehat{\mathbb{P}}_N^x$ y radio $\varepsilon \|x\|$.

A partir de lo anterior definimos el conjunto

$$\mathbb{X} := \left\{ x \in \mathbb{R}^m \;\middle|\; \sum_{i=1}^{m} x_i = 1,\ x_i \geq 0,\ \mathbb{E}_{\mathbb{Q}}[\zeta^x] \geq \mu \ \forall\, \mathbb{Q} \in \mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x) \right\}$$
$$= \left\{ x \in \mathbb{R}^m \;\middle|\; \sum_{i=1}^{m} x_i = 1,\ x_i \geq 0,\ \inf_{\mathbb{Q} \in \mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x)} \mathbb{E}_{\mathbb{Q}}[\zeta^x] \geq \mu \right\} \tag{15}$$

Entonces la versión Robusta Distribucional de (13) es

$$\widehat{J}_N := \underset{x \in \mathbb{X}}{\text{minimizar}} \quad \sup_{\mathbb{Q} \in \mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x)} \mathrm{Var}_{\mathbb{Q}}\left[\zeta^x\right]. \tag{16}$$

Llamaremos a este problema DR-MRK.

Pero por el Teorema 2 se tiene

$$\inf_{\mathbb{Q}\in\mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x)} \mathbb{E}_{\mathbb{Q}}[\zeta^x] = \frac{1}{N}\sum_{i=1}^{N}\widehat{\zeta}_i^x - \varepsilon\|x\|\,.$$

Lo que permite reescribir $\mathbb{X}$ como

$$\begin{aligned}\mathbb{X} &= \left\{ x\in\mathbb{R}^m \;\middle|\; \sum_{i=1}^{m} x_i = 1,\ x_i \geq 0,\ \frac{1}{N}\sum_{i=1}^{N}\widehat{\zeta}_i^x - \varepsilon\|x\| \geq \mu \right\} \\ &= \left\{ x\in\mathbb{R}^m \;\middle|\; \sum_{i=1}^{m} x_i = 1,\ x_i \geq 0,\ \frac{1}{N}\sum_{i=1}^{N}\langle x,\widehat{\xi}_i\rangle - \varepsilon\|x\| \geq \mu \right\}\end{aligned} \tag{17}$$

Por otro lado, introduciendo una variable de holgura se obtiene

$$\sup_{\mathbb{Q}\in\mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x)} \operatorname{Var}_{\mathbb{Q}}[\zeta^x] = \sup_{\substack{\eta\geq\mu,\\ \left(\eta-\frac{1}{N}\sum_{i=1}^N \widehat{\zeta}_i^x\right)^2\leq\varepsilon^2\|x\|^2}} \begin{cases} \sup\limits_{\mathbb{Q}\in\mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x)} & \operatorname{Var}_{\mathbb{Q}}[\zeta^x] \\ \text{sujeto a} & \mathbb{E}_{\mathbb{Q}}[\zeta^x]=\eta. \end{cases}$$

La primer tarea es reescribir el problema interno

$$\begin{cases} \sup\limits_{\mathbb{Q}\in\mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x)} & \operatorname{Var}_{\mathbb{Q}}[\zeta^x] \\ \text{sujeto a} & \mathbb{E}_{\mathbb{Q}}[\zeta^x]=\eta. \end{cases}$$

En ese sentido es importante el siguiente Teorema.

Teorema 6

Sea ζ un a variable aleatoria con distribución $\mathbb{P}$ y tal que se conoce su esperanza, es decir, se sabe que $\mathbb{E}_{\mathbb{P}}[\zeta] = \eta$, además, sea $\widehat{\zeta}_1, \ldots, \widehat{\zeta}_N$ una muestra de ζ y $\varepsilon^2 \geq \left(\frac{1}{N} \sum_{i=1}^{N} (\eta - \widehat{\zeta}_i) \right)^2$. Considerando $\mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)$ como la bola respecto a la métrica 2-Wasserstein de radio ε centrada en $\widehat{\mathbb{P}}_N$ la distribución empírica respecto a la muestra anterior, entonces

$$\left\{ \begin{array}{cc} \sup_{\mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} & \mathbb{E}_{\mathbb{Q}} \left[(\zeta - \eta)^2 \right] \\ \textit{sujeto a} & \mathbb{E}_{\mathbb{Q}} [\zeta] = \eta. \end{array} \right. = \left(\sqrt{\frac{1}{N} \sum_{i=1}^{N} \widehat{\zeta}_i^2 - \left(\frac{1}{N} \sum_{i=1}^{N} \widehat{\zeta}_i \right)^2} + \sqrt{\varepsilon^2 - \left(\frac{1}{N} \sum_{i=1}^{N} (\eta - \widehat{\zeta}_i) \right)^2} \right)^2 . \tag{18}$$

Por lo tanto

$$\left\{\begin{array}{ll} \sup & \mathrm{Var}_{\mathbb{Q}}\left[\zeta^x\right] \\ \mathbb{Q}\in\mathcal{B}_{\|x\|\varepsilon}(\widehat{\mathbb{P}}_N^x) & \\ \text{sujeto a} & \mathbb{E}_{\mathbb{Q}}\left[\zeta^x\right]=\eta. \end{array}\right. = \left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}\langle x,\widehat{\xi}_i\rangle^2-\frac{1}{N^2}\left(\sum_{i=1}^{N}\langle x,\widehat{\xi}_i\rangle\right)^2}+\sqrt{\varepsilon^2\|x\|^2-\left(\eta-\frac{1}{N}\sum_{i=1}^{N}\langle x,\widehat{\xi}_i\rangle\right)^2}\right)^2$$

Así pues, por este teorema se infiere que (16) es equivalente al problema de optimización:

$$\widehat{J}_N=\underset{x\in\mathbb{X}}{\text{minimizar}}\left\{\begin{array}{ll} \sup\limits_{\eta\geq\mu} & \left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}\langle x,\widehat{\xi}_i\rangle^2-\frac{1}{N^2}\left(\sum_{i=1}^{N}\langle x,\widehat{\xi}_i\rangle\right)^2}+\sqrt{\varepsilon^2\|x\|^2-\left(\mu-\frac{1}{N}\sum_{i=1}^{N}\langle x,\widehat{\xi}_i\rangle\right)^2}\right)^2 \\ \text{sujeto a} & \left(\eta-\dfrac{1}{N}\sum\limits_{i=1}^{N}\widehat{\zeta}_i^x\right)^2\leq\varepsilon^2\|x\|^2 \end{array}\right. \tag{19}$$

Pero el problema de maximización interno de (19) puede solucionarse explícitamente, en realidad dicho problema alcanza su valor óptimo en $\eta^* = \frac{1}{N}\sum_{i=1}^{N} \widehat{\zeta}_i^x$, por lo tanto, (19) se puede reescribir como

$$\widehat{J}_N = \underset{x\in\mathbb{X}}{\text{minimizar}} \left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}\langle x, \widehat{\xi}_i\rangle^2 - \frac{1}{N^2}\left(\sum_{i=1}^{N}\langle x, \widehat{\xi}_i\rangle\right)^2} + \varepsilon \|x\| \right)^2$$

$$= \begin{cases} \underset{x\in\mathbb{R}^m}{\text{minimizar}} & \left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}\langle x, \widehat{\xi}_i\rangle^2 - \frac{1}{N^2}\left(\sum_{i=1}^{N}\langle x, \widehat{\xi}_i\rangle\right)^2} + \varepsilon \|x\| \right)^2 \\ \text{sujeto a} & \sum_{i=1}^{N}\langle x, \widehat{\xi}_i\rangle - \varepsilon \|x\| \geq \mu, \\ & \sum_{i=1}^{m} x_i = 1. \end{cases} \tag{20}$$

El problema (20) se puede simplificar aun más.

Proposición 7

Sea M la matriz de tamaño $m \times N$ cuyas columnas son los vectores de la muestra $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ y sean $\mathbf{0}, \mathbf{e} \in \mathbb{R}^N$ los vectores columna de ceros y unos respectivamente A partir de estas convenciones se definen las matrices

$$E := \frac{1}{N} MM^T - \frac{1}{N^2}(M\mathbf{e})(M\mathbf{e})^T \quad y \quad L := \frac{1}{N}(Me)^T.$$

Ya que E es semidefinida positiva entoncessemidefinida positiva de modo que existe una matriz K tal que $E = KK^T$. Entonces (20) es equivalente al problema de optimización

$$\begin{cases} \inf\limits_{x \in \mathbb{R}^m} & \left(\left\|K^T x\right\| + \varepsilon \|x\|\right)^2 \\ \text{sujeto a} & Lx - \varepsilon \|x\| \geq \mu, \\ & e^T x = 1. \end{cases} \tag{21}$$

Se debe tener en cuenta que (21) puede no ser factible para algunos valores de ε, concretamente, dada L y μ se tiene que (21) es factible si

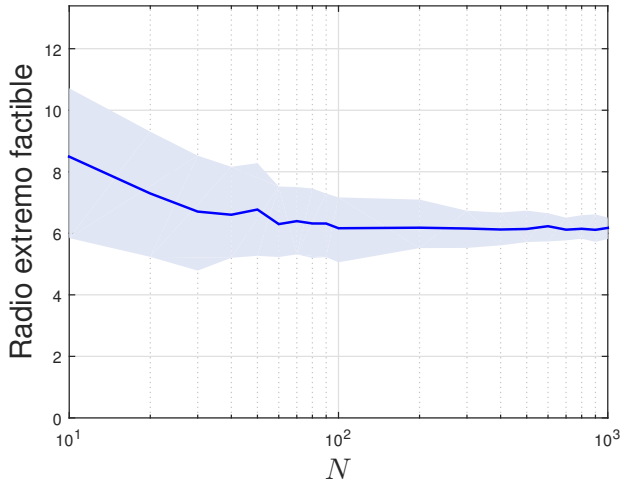
$$\varepsilon < \widehat{\epsilon}_N(\mu) := \begin{cases} \sup\limits_{x \in \mathbb{R}^m} & \dfrac{Lx - \mu}{\|x\|} \\ \text{sujeto a} & \sum\limits_{i=1}^{m} x_i = 1. \end{cases}$$

La dependencia de N en $\widehat{\epsilon}_N(\mu)$ se debe a que L depende de la muestra. En adelante llamaremos *radio extremo factible* a la expresión $\widehat{\epsilon}_N(\mu)$.

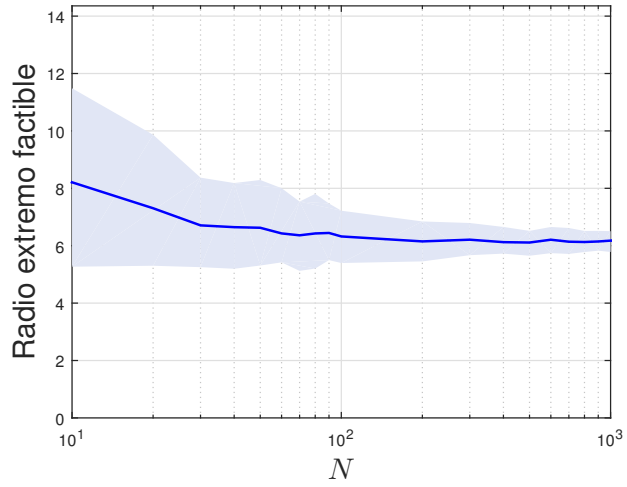
Algunos resultados numéricos

Los resultados numéricos que se traducen en las siguiente gráficas originadas por simulaciones realizadas para un portafolio compuesto de cuatro bienes, es decir, $m = 4$, la distribución de ξ es multinormal con matriz de covarianza C y vector de medias $\mathbf{m}$ dados por

$$C = \begin{bmatrix} 185 & 86{,}5 & 80 & 20 \\ 86{,}5 & 196 & 76 & 13{,}5 \\ 80 & 76 & 411 & -19 \\ 20 & 13{,}5 & -19 & 25 \end{bmatrix} \quad \text{y} \quad \mathbf{m} = (14, 12, 15, 17).$$



(a)



(b)

Figura: Radio extremo factible $\widehat{\epsilon}_N(\mu)$ como función de N. En (a) para $\mu = 20$ y (b) $\mu = 40$.

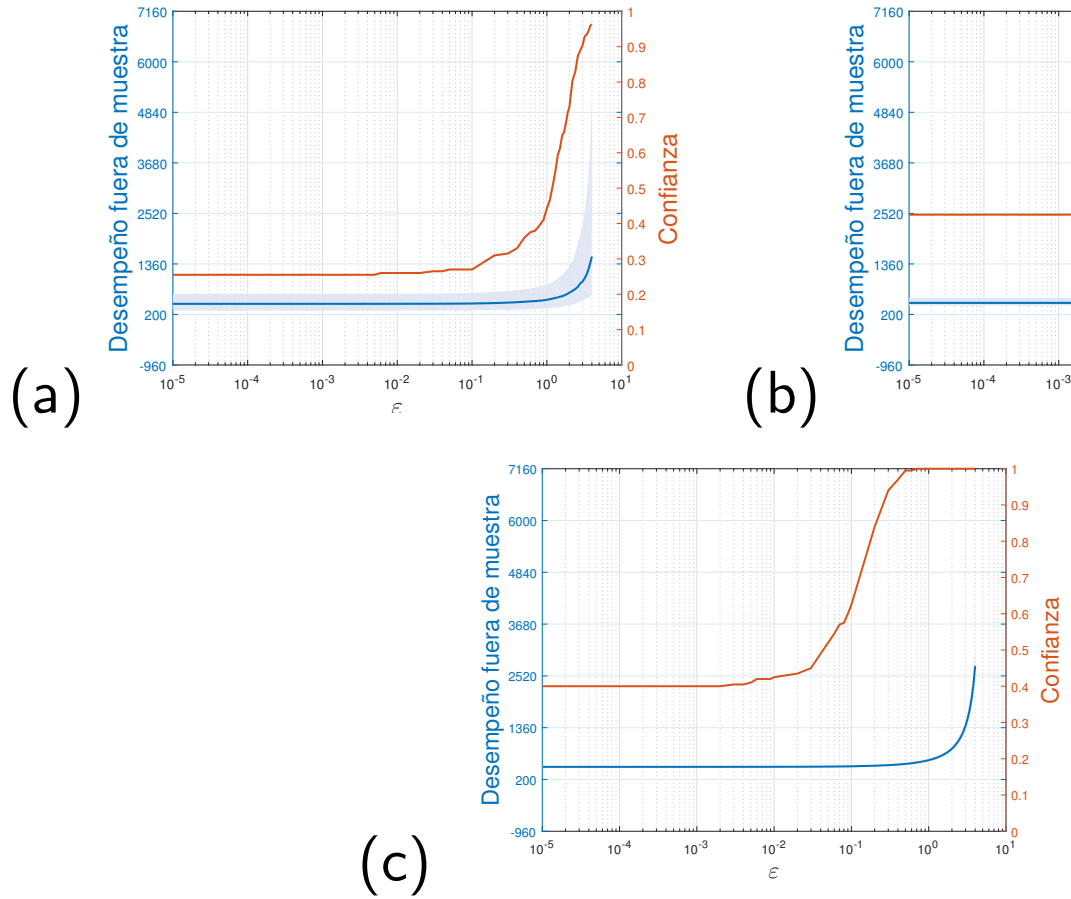


Figura: Desempeño fuera de muestra $\mathrm{Var}_{\mathbb{P}}[\langle \widehat{x}_N(\varepsilon)\rangle]$ (eje izquierdo, linea azul y área sombreada) y confianza $\mathbb{P}^N\left[\mathrm{Var}_{\mathbb{P}}[\langle \widehat{x}_N(\varepsilon)\rangle] \leq \widehat{J}_N(\varepsilon)\right]$ (linea naranja) como función de ε. En (a) para $N = 30$, (b) $N = 300$ y (c) $N = 3000$. En todos los caso $\mu = 20$

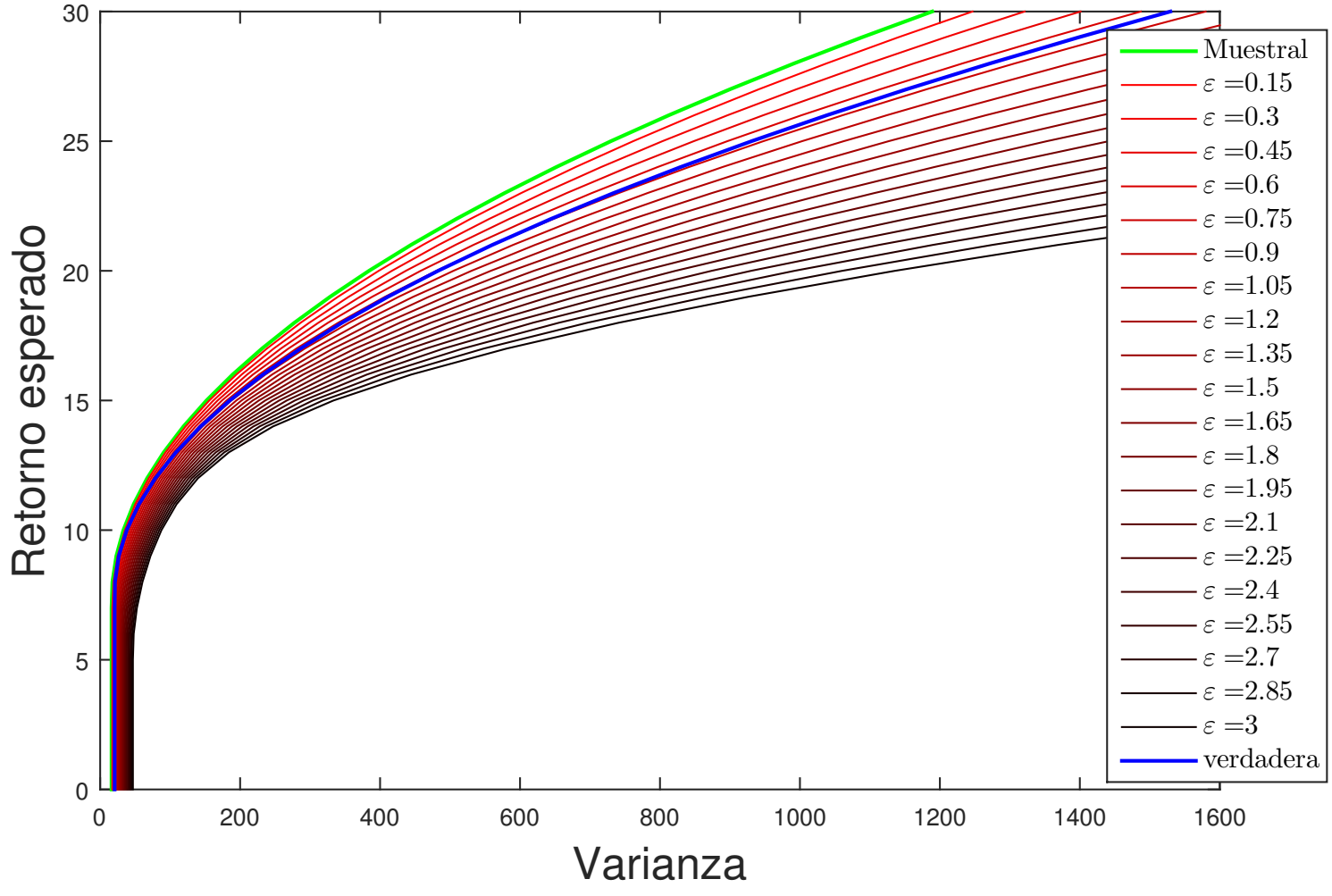


Figura: Curvas de frontera eficiente estimada por Wasserstein para varios valores de ε con una muestra de tamaño $N = 100$.

Determinación del parámetro bandwidth en le estimación de funciones de densidad por kernels

Sea ξ una variable aleatoria y f su función de densidad, dada una muestra $\widehat{\xi}_1, \ldots, \widehat{\xi}_N$ este método de estimación propone el estimador de f dado por

$$\widehat{f}_h(x) := \frac{1}{Nh} \sum_{i=1}^{N} \mathcal{K}\left(\frac{x - \widehat{\xi}_i}{h}\right).$$

Este estimador depende del parámetro $h > 0$ comúnmente conocido como *ancho de banda* o *parámetro de suavidad*, también depende de $\mathcal{K}$ la cual es una función de densidad simétrica al rededor de cero y tal que $\int x^2 \mathcal{K}(x) dx = 1$. Frecuentemente se considera $\mathcal{K}$ como la función de densidad normal estándar, no obstante $\mathcal{K}$ puede ser cualquier función de densidad que satisfaga las condiciones de simetría y varianza.

El método consiste en encontrar un $h > 0$ tal que minimice el *error cuadrático integrado medio* (mean integrated squared error)

$$\mathrm{MISE}(h) := \mathbb{E}_{\mathbb{P}^N}\left[\int \left(\widehat{f}_h(x) - f(x)\right)^2 dx\right]. \tag{22}$$

La aleatoriedad en esta expresión esta en el vector aleatorio $\left(\widehat{\xi}_1, \ldots, \widehat{\xi}_N\right)$ cuya distribución es $\mathbb{P}^N = \mathbb{P} \times \cdots \times \mathbb{P}$, donde $\mathbb{P}$ es la distribución inducida por la función de densidad f, de modo que $\mathbb{P}$ también es desconocida.

Por la linealidad del valor esperado la expresión (22) se pude reformular de la siguiente manera:

$$\mathrm{MISE}(h) = \mathbb{E}_{\mathbb{P}^N}\left[\int \left(\widehat{f}_h(x)\right)^2 dx\right] - 2\mathbb{E}_{\mathbb{P}^N}\left[\int \widehat{f}_h(x) f(x) dx\right] + \int (f(x))^2 dx.$$

Considerando J la parte que depende de h en la expresión anterior, es decir

$$J(h) := \mathbb{E}_{\mathbb{P}^N}\left[\int \left(\widehat{f}_h(x)\right)^2 dx\right] - 2\mathbb{E}_{\mathbb{P}^N}\left[\int \widehat{f}_h(x) f(x) dx\right], \quad (23)$$

entonces se infiere que h minimiza la expresión MISE si y solo si minimiza J, de modo que enfocamos nuestra atención en minimizar J, para tal fin intentaremos reescribir J de tal manera que minimizar J sea un problema de optimización estocástica.

Por un lado se observa que

$$
\begin{aligned}
\mathbb{E}_{\mathbb{P}^N}\left[\int\left(\widehat{f}_h(x)\right)^2 dx\right] &= \int\cdots\int\int\left(\widehat{f}_h(x)\right)^2 dx\mathbb{P}(d\xi_1)\cdots\mathbb{P}(d\xi_N)\\
&= \int\cdots\int\int\tfrac{1}{N^2h^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathcal{K}\left(\tfrac{x-\xi_i}{h}\right)\mathcal{K}\left(\tfrac{x-\xi_j}{h}\right)dx\mathbb{P}(d\xi_1)\cdots\mathbb{P}(d\xi_N)\\
&= \tfrac{1}{N^2h^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\int\cdots\int\int\mathcal{K}\left(\tfrac{x-\xi_i}{h}\right)\left(\tfrac{x-\xi_j}{h}\right)dx\mathbb{P}(d\xi_1)\cdots\mathbb{P}(d\xi_N)\\
&= \tfrac{1}{N^2h^2}\int\int\left(\mathcal{K}\left(\tfrac{x-\xi}{h}\right)\right)^2 dx\mathbb{P}(d\xi)\\
&\quad +\tfrac{N-1}{Nh^2}\int\int\int\mathcal{K}\left(\tfrac{x-\xi}{h}\right)\mathcal{K}\left(\tfrac{x-\zeta}{h}\right)dx\mathbb{P}(d\xi)\mathbb{P}(d\zeta)\\
&= \tfrac{1}{N^2h^2}\mathbb{E}_{\mathbb{P}\sim\xi}\left[\int\left(\mathcal{K}\left(\tfrac{x-\xi}{h}\right)\right)^2 dx\right]+\tfrac{N-1}{Nh^2}\mathbb{E}_{\mathbb{P}\times\mathbb{P}\sim(\xi,\zeta)}\left[\int\mathcal{K}\left(\tfrac{x-\xi}{h}\right)\mathcal{K}\left(\tfrac{x-\zeta}{h}\right)dx\right]\\
&= \mathbb{E}_{\mathbb{P}\times\mathbb{P}\sim(\xi,\zeta)}\left[\tfrac{1}{N^2h^2}\int\left(\mathcal{K}\left(\tfrac{x-\xi}{h}\right)\right)^2 dx+\tfrac{N-1}{Nh^2}\int\mathcal{K}\left(\tfrac{x-\xi}{h}\right)\mathcal{K}\left(\tfrac{x-\zeta}{h}\right)dx\right].
\end{aligned}
$$

Por otro lado, se tiene

$$
\begin{aligned}
\mathbb{E}_{\mathbb{P}^N}\left[\mathbb{E}_{\mathbb{P}}\left[\widehat{f}_h(x)\right]\right] &= \int\cdots\int\int \widehat{f}_h(x) f(x) dx \mathbb{P}(\xi_1)\cdots\mathbb{P}(\xi_N) \\
&= \int\cdots\int\int \frac{1}{Nh}\sum_{i=1}^{N} \mathcal{K}\left(\frac{x-\xi_i}{h}\right) f(x) dx \mathbb{P}(\xi_1)\cdots\mathbb{P}(\xi_N) \\
&= \frac{1}{Nh}\sum_{i=1}^{N}\int\cdots\int\int \mathcal{K}\left(\frac{x-\xi_i}{h}\right) f(x) dx \mathbb{P}(\xi_1)\cdots\mathbb{P}(\xi_N) \\
&= \frac{1}{Nh}\sum_{i=1}^{N}\int\int \mathcal{K}\left(\frac{x-\xi_i}{h}\right) f(x) dx \mathbb{P}(\xi_i) \\
&= \frac{1}{h}\int\int \mathcal{K}\left(\frac{x-\xi}{h}\right) f(x) dx \mathbb{P}(d\xi) \\
&= \mathbb{E}_{\mathbb{P}\times\mathbb{P}\sim(\xi,\zeta)}\left[\frac{1}{h}\mathcal{K}\left(\frac{\zeta-\xi}{h}\right)\right].
\end{aligned}
$$

Por lo tanto, la expresión que define J dada en (23) es igual a

$$J(h) = \mathbb{E}_{\mathbb{P}\times\mathbb{P}\sim(\xi,\zeta)}\left[\frac{1}{N^2h^2}\int\left(\mathcal{K}\left(\frac{x-\xi}{h}\right)\right)^2 dx + \frac{N-1}{Nh^2}\int\mathcal{K}\left(\frac{x-\xi}{h}\right)\mathcal{K}\left(\frac{x-\zeta}{h}\right)dx - \frac{2}{h}\mathcal{K}\left(\frac{\zeta-\xi}{h}\right)\right]. \quad (24)$$

Para efectos de notación definimos

$$F(h,\xi,\zeta) := \frac{1}{Nh^2}\int\left(\mathcal{K}\left(\frac{x-\xi}{h}\right)\right)^2 dx + \frac{N-1}{Nh^2}\int\mathcal{K}\left(\frac{x-\xi}{h}\right)\mathcal{K}\left(\frac{x-\zeta}{h}\right)dx - \frac{2}{h}\mathcal{K}\left(\frac{\zeta-\xi}{h}\right).$$

De modo que minimizar $J(h)$ es en realidad un problema de optimización estocástica expresado como

$$J^* := \min_{h\geq 0} J(h) = \min_{h\geq 0}\mathbb{E}_{\mathbb{P}\times\mathbb{P}\sim(\xi,\zeta)}\left[F(h,\xi,\zeta)\right] \quad (25)$$

donde $\mathbb{P}$ es desconocida.

Caso kerner normal estándar:

En este caso asumimos que $\mathcal{K}$ tiene la siguiente forma:

$$\mathcal{K}(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}.$$

Por lo tanto, para este caso se tiene

$$F(h, \xi, \zeta) := \frac{1}{2\sqrt{\pi} N h^3} + \frac{N-1}{2\sqrt{\pi} N h^3} e^{-\frac{(\xi-\zeta)^2}{4h^2}} - \frac{2}{\sqrt{2\pi} h} e^{-\frac{(\xi-\zeta)^2}{2h^2}}.$$

Para la versión robusta consideramos $\Xi = \mathbb{R}^2$ y la métrica 2-Wasserstein con función de costo d como la distancia euclidiana en $\mathbb{R}^2$. Así pues, asumimos que el tamaño de la muestra N es par, es decir, $N = 2M$, entonces $(\widehat{\xi}_1, \widehat{\xi}_{N+1}), \ldots, (\widehat{\xi}_N, \widehat{\xi}_{2N})$ es una muestra de tamaño N de (ξ, ζ) lo que permite definir

$$\widehat{\mathbb{P}}_N = \sum_{i=1}^{N} \delta_{(\widehat{\xi}_i, \widehat{\xi}_{N+i})}.$$

Por lo tanto, tenemos que la contraparte robusta distribucional de (25) es

$$\widehat{J}_N = \min_{h \geq 0} \sup_{\mathbb{Q} \in \mathcal{B}_\varepsilon(\widehat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}} \left[F(h, \xi, \zeta) \right]$$

Por el Teorema 2 este último problema es equivalente a

$$\begin{cases} \inf\limits_{h,\lambda,s} & \lambda \varepsilon^p + \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{(\xi,\zeta) \in \mathbb{R}^2} \left(F(h, \xi, \zeta) - \lambda \left\| (\xi, \zeta) - (\widehat{\xi}_i, \widehat{\xi}_{N+i}) \right\|^2 \right) \leq s_i \quad \forall i = 1, \ldots, N, \\ & \lambda \geq 0, \\ & h \geq 0. \end{cases} \tag{26}$$

Analizando la función que determina la primera restricción se sigue que el problema anterios es equivalente a:

$$\begin{cases} \inf\limits_{h,\lambda,s} & \lambda\varepsilon^p + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} s_i \\ \text{sujeto a} & \sup\limits_{\xi\in\mathbb{R}} \left(F(h,\xi,-\xi+\widehat{\xi}_i+\widehat{\xi}_{N+i}) - 2\lambda(\xi-\widehat{\xi}_i)^2 \right) \leq s_i \quad \forall i = 1,\ldots,N, \\ & \lambda \geq 0, \\ & h \geq 0. \end{cases} \tag{27}$$

La idea para solucionar el problema anterior es implementar un algoritmo cutting planes. (Work in progress)

Gracias por su atención.

Referencias I

Esfahani, PM. y Kuhn, D. A.
Data-driven Distributionally Robust Optimization Using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations.
arXiv preprint arXiv:1505.05116v2, 2016.

Lee, C. y Mehrotra, S.
A distributionally-Robust Optimization approach for finding support vector machines.
Optimization Online, 2015.

Pflug, G. y Wozabal, D.
Ambiguity in portfolio selection.
Quantitative finance, 435-442, 2006.