



UNIVERSIDAD NACIONAL DE COLOMBIA

Comparación de Máquinas de Soporte Vectorial vs. Regresión Logística. ¿Cuál es más recomendable para discriminar?

Diego Alejandro Salazar Blandón

Universidad Nacional de Colombia

Facultad de Ciencias

Escuela de Estadística

Medellín, Colombia

2012

Comparación de Máquinas de Soporte Vectorial vs. Regresión Logística. ¿Cuál es más recomendable para discriminar?

Diego Alejandro Salazar Blandón

Tesis de grado presentada como requisito parcial para optar al título de:
Magister en Ciencias-Estadística

Director:

Juan Carlos Salazar Uribe, Ph.D.

Línea de Investigación:

Bioestadística

Grupos de Investigación:

Neurociencias de la Universidad Nacional de Colombia, Sede Bogotá.

Investigación en Estadística de la Universidad Nacional de Colombia, Sede Medellín.

Universidad Nacional de Colombia

Facultad de Ciencias

Escuela de Estadística

Medellín, Colombia

2012

*A la memoria de mi amigo, ejemplo y abuelo.
Benjamin Antonio Blandón Berrio (1921-2004)*

Agradecimientos

A Dios por brindarme unos maravillosos padres y familiares. Sin su comprensión y apoyo incondicional no hubiera sido posible lograr esta meta.

Al profesor Juan Carlos Salazar Uribe, por incentivar y creer en mi trabajo.

A Jorge Iván Vélez, cuyas observaciones y sugerencias fueron fundamentales en el desarrollo de esta tesis.

Resumen

La clasificación de objetos es un problema muy común en el trabajo estadístico aplicado. Si se tiene un conjunto de datos X correspondientes a una muestra de una población en el que cada uno de sus elementos pertenece a una de dos clases, el objetivo de los métodos de clasificación es determinar a cuál de esas dos clases pertenecerá una *nueva* observación. Uno de los métodos más utilizados es la regresión logística (RL); su validez y desempeño han sido ampliamente demostrados en la literatura. Recientemente, las Máquinas de Soporte Vectorial (SVM), un método alternativo basado en procesos algorítmicos, proporciona un enfoque diferente a la solución de este problema. En este trabajo se exponen los principios básicos de RL y SVM y se comparan, vía simulación, para dar respuesta a la pregunta de cuál es más recomendable para discriminar cuando la población puede clasificarse en dos categorías. Finalmente se presentan dos aplicaciones con datos provenientes de microarreglos en los que se midieron los niveles de expresión de genes en pacientes con diabetes y enfermedad de Alzheimer.

Palabras clave: Máquinas de Soporte Vectorial, Regresión Logística, Clasificación, Simulación Estadística, Genética.

Abstract

The classification of individuals or objects is a common problem in applied statistics. For instance, if X is a data set corresponding to a sample from a specific population in which all its observations belong to one of two categories, the goal of classification methods is to decide to which class a *new* observation will be in. One of the most and widely used classification methods is logistic regression (LR); its properties and performance have been extensively studied in the literature. Recently, Support Vector Machine (SVM), an alternative method based on highly structured algorithms, has provided a new solution to the classification problem. In this work, the fundamentals of LR and SVM are described. Also, using statistical simulation, we address the question of which of them is better to discriminate when the population can be classified in two categories. Finally, two applications

with real data from microarray experiments in diabetes and Alzheimer's disease are presented as illustration.

Keywords: Support Vector Machines, Logistic Regression, Classification, Statistical Simulation, Genetics.

Índice General

Agradecimientos	IV
Resumen	v
1. Introducción	2
2. Métodos de Clasificación	6
2.1. SVM para dos Grupos	6
2.2. Regresión Logística	10
3. Estrategias de Comparación	12
3.1. Escenarios Univariados	12
3.2. Escenarios Normales Multivariantes	13
3.3. Estrategia de Simulación	14
4. Resultados de simulaciones	16
4.1. Univariados	16
4.1.1. Distribución Normal	17
4.1.2. Distribución Poisson	19
4.1.3. Distribución Exponencial	21
4.1.4. Distribución Cauchy	23
4.1.5. Distribución Lognormal	25
4.2. Combinación de Distribuciones	27
4.2.1. Distribución Cauchy-Normal	27
4.2.2. Distribución Normal-Poisson	29
4.2.3. Distribución Normal-Exponencial	31

4.3. Distribuciones Multivariadas	33
4.3.1. Distribución Normal Bivariada	33
4.3.2. Distribución Normal Multivariada con $p = 200$	33
5. Aplicaciones Genéticas	46
5.1. Expresión Genética	46
5.2. Datos sobre Diabetes	48
5.3. Datos sobre Alzheimer	52
6. Conclusiones Generales	55
A. Anexo: Programas en R	57
B. Anexo: Algoritmo Aplicaciones	60
C. Anexo: Resultados Multivariados Adicionales	64
C.1. Normal Multivariada ($p = 10$)	65
C.2. Normal Multivariada ($p = 20$)	68
C.3. Normal Multivariada ($p = 50$)	71

Índice de Figuras

2.1. Ejemplo ilustrativo SVM	7
2.2. Ejemplo aplicación SVM	9
4.1. Distribución Normal	17
4.2. Distribución Poisson	19
4.3. Distribución Exponencial	21
4.4. Distribución Cauchy	23
4.5. Distribución Lognormal	25
4.6. Distribución Cauchy-Normal	27
4.7. Distribución Normal-Poisson	29
4.8. Distribución Normal-Exponencial	31
4.9. Distribución Normal Bivariada $\Sigma_1 = \Sigma_2$	34
4.10. Distribución Normal Bivariada $\Sigma_1 = 2\Sigma_2$ diferente	35
4.11. Distribución Normal Bivariada $\Sigma_1 = 3\Sigma_2$	38
4.12. Distribución Normal Multivariada ($p = 200$) $\Sigma_1 = \Sigma_2$	40
4.13. Distribución Normal Multivariada ($p = 200$) $\Sigma_1 = 2\Sigma_2$	42
4.14. Distribución Normal Multivariada ($p = 200$) $\Sigma_1 = 3\Sigma_2$	44
5.1. Procedimiento general para obtención de GE.	47
5.2. Matriz de dispersión niveles de expresión de genes en DT2	50
5.3. Genes vs. MCR (Diabetes)	51
5.4. Matriz de dispersión niveles de expresión de genes en Alzheimer	53
5.5. Genes vs. MCR (Alzheimer)	54
C.1. Distribución Normal Multivariada ($p = 10$) $\Sigma_1 = \Sigma_2$	65
C.2. Distribución Normal Multivariada ($p = 10$) $\Sigma_1 = 2\Sigma_2$	66

C.3. Distribución Normal Multivariada ($p = 10$) $\Sigma_1 = 3\Sigma_2$	67
C.4. Distribución Normal Multivariada ($p = 20$) $\Sigma_1 = \Sigma_2$	68
C.5. Distribución Normal Multivariada ($p = 20$) $\Sigma_1 = 2\Sigma_2$	69
C.6. Distribución Normal Multivariada ($p = 20$) $\Sigma_1 = 3\Sigma_2$	70
C.7. Distribución Normal Multivariada ($p = 50$) $\Sigma_1 = \Sigma_2$	71
C.8. Distribución Normal Multivariada ($p = 50$) $\Sigma_1 = 2\Sigma_2$	72
C.9. Distribución Normal Multivariada ($p = 50$) $\Sigma_1 = 3\Sigma_2$	73

Índice de Tablas

2.1. Kernels más utilizados en SVM.	9
3.1. Distribuciones de probabilidad univariadas consideradas	13
4.1. Resultados distribución Normal.	18
4.2. Resultados distribución Poisson.	20
4.3. Resultados distribución Exponencial.	22
4.4. Resultados distribución Cauchy.	24
4.5. Resultados distribución Lognormal.	26
4.6. Resultados distribución Cauchy-Normal.	28
4.7. Resultados distribución Normal-Poisson.	30
4.8. Resultados distribución Normal-Exponencial.	32
4.9. Resultados distribución Normal Bivariada $\Sigma_1 = \Sigma_2$	36
4.10. Resultados distribución normal bivariada $\Sigma_1 = 2\Sigma_2$	37
4.11. Resultados distribución Normal Bivariada $\Sigma_1 = 3\Sigma_2$	39
4.12. Resultados distribución Normal Multivariada ($p = 200$) $\Sigma_1 = \Sigma_2$	41
4.13. Resultados distribución Normal Multivariada ($p = 200$) $\Sigma_1 = 2\Sigma_2$	43
4.14. Resultados distribución Normal Multivariada ($p = 200$) $\Sigma_1 = 3\Sigma_2$	45
5.1. Resumen estadístico genes (Diabetes)	49
5.2. Resumen estadístico genes (Alzheimer)	52

1. Introducción

Si se establecen grupos específicos dentro de una población, los métodos de clasificación buscan la construcción de una función que, a partir de una muestra de una población, permita discriminar nuevos elementos, es decir, se busca que esta función pueda pronosticar de una manera óptima a cuál grupo pertenece una nueva observación (Anderson 1984). En las áreas de la salud, por ejemplo, se está interesado en contar con funciones de este tipo que permitan establecer la presencia de determinada enfermedad mediante el estudio de diagnósticos previos de los sujetos de una población (Lu et al. 2003), o establecer el riesgo de poseerla mediante el análisis de los genotipos particulares de cada individuo (Dubey & Realff 2004).

Los métodos estadísticos tradicionales como el análisis discriminante lineal introducido por Fisher (1936), exhiben soluciones óptimas para el problema de clasificación en la medida en que los supuestos sobre los que se basan se satisfagan. Sin embargo, se presentan grandes inconvenientes cuando en aplicaciones reales estos supuestos teóricos no pueden ser validados. En estudios genéticos con familias o dentro de un grupo étnico específico, Whittemore (2004) discute el supuesto de correlación entre grupos en estudios caso/control y propone el estudio de familias nucleares usando regresión logística. Por tanto, ante un inconveniente en el cumplimiento de los supuestos, se pone en duda la veracidad de los resultados obtenidos con las implicaciones que una mala clasificación puede llegar a tener. Por ejemplo, podría ocurrir que dado un diagnóstico equivocado de cáncer, se someta a una persona sana a un tratamiento de radiación o consumo de químicos causando el deterioro de la salud ó incluso su muerte.

Conscientes de estas falencias, se han hecho esfuerzos en encontrar vías alternas que debilitan los supuestos de los modelos clásicos, entre ellos la regresión logística (RL), propuesta y estudiada por Cornfield (1962), Cox (1966), Day & Kerridge (1967) y Hosmer & Lemeshow (1989). Dados sus escasos supuestos, la RL es una de las técnicas que más se encuentra en la literatura para dar solución a los problemas de clasificación. Por otro lado, los avances tecnológicos en el área de la computación de las últimas décadas, han promovido el desarrollo de nuevas metodologías basadas en cálculos iterativos ó algoritmos como las redes neuronales (Neural Networks, en inglés) ó las máquinas de aprendizaje (Machine Learning, en inglés). Sin embargo, un enfoque puramente computacional trae consigo sus desventajas, porque si bien en su utilización en la mayoría de los casos se garantizan resultados, estos métodos se pueden llegar a convertir en un proceso en el que se arrojan datos en una “caja negra” que devuelve soluciones sin que el usuario sepa qué ocurre dentro, limitado la interpretación de las soluciones y poniendo en duda, nuevamente, los resultados obtenidos. Lo ideal desde este punto de vista es hallar un equilibrio entre la sustentación teórica y el uso de los algoritmos computacionales. Las Máquinas de Soporte Vectorial (Cortes & Vapnik 1995) ó SVM por su nombre en inglés (Support Vector Machines), son un método de clasificación que combina el uso de la computación con la argumentación teórica. Estas características han dotado a las SMV de una gran reputación y han promovido su implementación en diferentes áreas (Hongdong et al. 2009, Tripathi et al. 2006, Crisler et al. 2008).

No obstante, ajustar un enfoque teórico y computacional, visto desde otra perspectiva, implica también la combinación de las desventajas de uno u otro enfoque. Se tienen ya dos interrogantes: ¿son exigentes los supuestos de la parte teórica? y ¿qué sucede dentro del algoritmo?. En este trabajo se busca dar respuesta a estos interrogantes dentro del contexto de las SVM haciendo uso de las bondades computacionales con las que se cuenta en la actualidad. Para ello se realiza un estudio de simulación que involucre diferentes escenarios, similares a los presentados en Hernández & Correa (2009), para comparar la efectividad de las SVM y RL.

Desde la aparición de las SVM en los años noventa han surgido algunas propuestas

para compararlas con métodos ya existentes, entre ellos RL. Sin embargo, en la mayoría de los casos se usan datos reales que limitan su comparación. Algunos ejemplos incluyen la comparación de 20 métodos de clasificación utilizando datos provenientes de microarreglos y entre los que se encuentran las SVM y RL (Lee et al. 2005), y un estudio con datos de mortalidad hospitalaria en pacientes en estado crítico debido a neoplasias malignas hematológicas (Verplancke et al. 2008). En ambos casos, al comparar la RL con las SVM los autores encontraron que no existía diferencia significativa entre los dos métodos. Sin embargo, las SVM requirieron menos variables que la RL para lograr una tasa de clasificación errónea equivalente a la arrojada por la RL. Por otro lado, Shou et al. (2009) comparan las SVM, RL y las redes neuronales para el diagnóstico de tumores benignos con base a imágenes a tres tipos de potencias diferentes. Los autores concluyen que no hay diferencia estadística significativa entre los tres métodos. Similarmente, Westreich et al. (2010) presenta las redes neuronales y las SVM como una alternativa a RL y realizan un estudio comparativo entre los supuestos de cada modelo y la posibilidad de implementarlos en paquetes estadísticos reconocidos como R, SAS y Stata. Los autores concluyen que para compararlos es necesario un arduo trabajo de simulación.

Concretamente, lo que se propone en esta investigación es realizar un estudio de simulación donde se pueda poner a prueba el comportamiento de las SVM y RL donde se controlen y varíen los parámetros en los datos de entrenamiento. Estos parámetros incluyen la cantidad de individuos y variables, la correlación entre variables y las distribuciones muestrales, entre otros. Esto permitirá generar datos con ciertas condiciones y finalmente decidir en qué caso es más recomendable usar uno u otro método de clasificación, lo cual constituye el aporte más importante de este trabajo. Se comenzará haciendo un desarrollo teórico de los fundamentos de cada método para luego exponer la estrategia de comparación, los algoritmos, los resultados obtenidos, dos aplicaciones con datos reales provenientes de experimentos con microarreglos en diabetes tipo 2 y enfermedad de Alzheimer, y finalmente una discusión general sobre las conclusiones obtenidas y algunas recomendaciones.

Esta tesis está organizada como sigue. En el capítulo 2 se describen aspectos teóricos relacionados con SVM y RL. En el capítulo 3 se presenta la estrategia de simulación. El capítulo 4 presentamos los resultados de las simulaciones. En el capítulo 5 se incluyen dos aplicaciones y en el capítulo final se presentan las conclusiones, recomendaciones y posibles direcciones futuras del estudio.

2. Métodos de Clasificación

En este capítulo se exponen los principios teóricos en los que basan cada uno de los métodos de clasificación. Sin entrar en detalles formales, se presentan las ideas principales detrás de cada metodología que permitan entender su funcionamiento.

2.1. SVM para dos Grupos

Las SVM aparecen en los años noventa como un método de clasificación óptimo, está constituido por un conjunto de algoritmos de aprendizaje supervisado desarrollados por Cortes & Vapnik (1995) junto con su equipo de los laboratorios AT&T. El nombre SVM fue explícitamente usado por primera vez por Cortes & Vapnik (1995). Su propuesta es la unión de dos ideas que ya habían aparecido individualmente en años anteriores:

1. El uso de los *kernels* y su interpretación geométrica, introducida por Aizerman et al. (1964)
2. La construcción de un hiperplano de separación óptimo en un contexto no paramétrico, desarrollado por Vapnik & Chervonenkis (1969).

En Moguerza & Muñoz (2006) y Tibshirani & Friedman (2008) se considera un problema de clasificación donde la función discriminante es no lineal (ver figura 2.1a) y se supone la existencia de un mapeo ó función kernel Φ a un “espacio característico” en el que los datos son linealmente separables (ver figura 2.1b). En este nuevo espacio, cada dato de la muestra es considerado como un punto de un espacio p -dimensional, donde p es el número de variables en el conjunto de datos.

Al aplicar Φ a los datos originales se obtiene una nueva muestra $\{(\Phi(\mathbf{x}_i), y_i)\}_{i=1}^n$ donde $y_i = \{-1, 1\}$ indica los dos posibles categorías (o clase) a las que pertenece

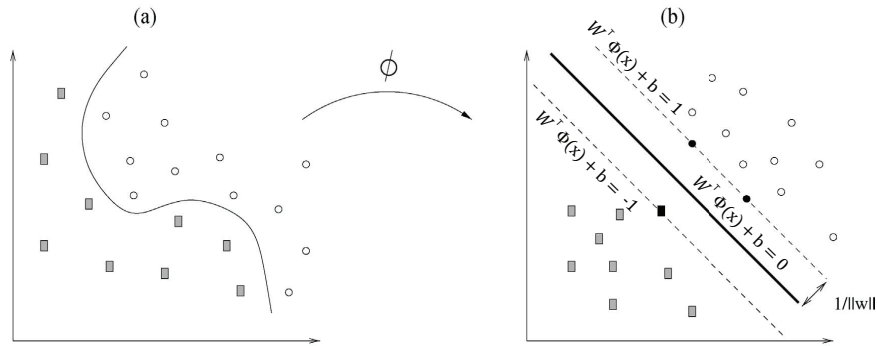


Figura 2.1.: Una ilustración de un modelo de SVM para dos grupos modificado de Moguerza & Muñoz (2006). El panel (a) muestra los datos y una función discriminante no lineal. En (b) se presentan los datos después aplicar la función kernel Φ .

cada dato, de tal forma que cualquier hiperplano de separación que equidista al punto más cercano de cada clase (ver puntos en negros en la figura 2.1b) se denota por $\mathbf{w}^T \Phi(\mathbf{x}) + b = 0$. Bajo el supuesto de separabilidad dado por el Teorema de Cover (Cover 1965), se pueden manipular \mathbf{w} y b de tal forma que $|\mathbf{w}^T \Phi(\mathbf{x}) + b| = 1$ para los puntos más cercanos al hiperplano en cada categoría. De esta forma se garantiza

$$\mathbf{w}^T \Phi(\mathbf{x}) + b \begin{cases} \geq 1, & \text{si } y_i = 1 \\ \leq -1, & \text{si } y_i = -1 \end{cases}.$$

para cada $i \in 1, \dots, n$.

La distancia del punto más cercano de cada clase al hiperplano es $1/\|w\|$ y la distancia entre los dos grupos es $2/\|w\|$. Tal distancia se conoce como *borde* o *margen*. Maximizar el margen implica resolver

$$\min_{\mathbf{w}, b} \|\mathbf{w}\|^2 \quad (2.1)$$

sujeto a $y_i(\mathbf{w}^T\Phi(\mathbf{x}) + b) \geq 1$ para $i = 1, \dots, n$.

Si \mathbf{w}^* y b^* son la solución de (2.1), estos determinan un hiperplano en el espacio característico

$$D^*(\mathbf{x}) = (\mathbf{w}^*)^T\Phi(\mathbf{x}) + b^* = 0.$$

Los puntos $\Phi(x_i)$ que satisfacen $y_i((\mathbf{w}^*)^T\Phi(\mathbf{x}) + b^*) = 1$ son llamados *vectores soporte* y de ellos depende la solución del problema de optimización. Las SVM son entonces, una serie de algoritmos computacionales que ayudan a resolver este problema de clasificación. Dentro de estos algoritmos es natural pensar que dependerán de la elección de la función Φ que transforma la muestra original, lo cual es cierto pero no directamente. Para poder encontrar \mathbf{w}^* y b^* es importante conocer el *producto interno* $\Phi(\mathbf{x}_i)^T \cdot \Phi(\mathbf{x}_j)$, asociado al nuevo espacio característico. La función definida por el *producto interno* es conocida como núcleo ó kernel, por lo que sólo es necesario conocer el kernel para determinar Φ y su *producto interno*. Los kernel más usados en la SVM (Karatzoglou et al. 2006) se presentan en la tabla 2.1.

Una vez encontrado el hiperplano de margen óptimo en el nuevo espacio, este se proyecta en el espacio original de los datos obteniéndose una función discriminante. Por ejemplo en la figura 2.2a se muestran algunos datos en \mathbb{R}^2 donde se evidencian dos grupos caracterizados por puntos negros y blancos (casos y controles, respectivamente) que no son linealmente separables. Luego, en la figura 2.2b, mediante un mapeo, los datos se llevan a \mathbb{R}^3 donde son separables por un plano de tal manera que al proyectarlo en el espacio original se tiene una función discriminante circular.

Tabla 2.1.: Kernels más utilizados en SVM.

Kernel	Función
Lineal	$\mathbf{x}_i^T \mathbf{x}_j$
Polinomial	$(\mathbf{x}_i^T \mathbf{x}_j + 1)^q$, q es el grado del polinomio
Gaussiano	$e^{-\frac{\ \mathbf{x}_i - \mathbf{x}_j\ ^2}{2\sigma^2}}$
Radial	$e^{-\gamma \ \mathbf{x}_i - \mathbf{x}_j\ ^2}$, $\gamma \geq 0$
Radial Laplace	$e^{-\gamma \ \mathbf{x}_i - \mathbf{x}_j\ }$, $\gamma \geq 0$
ANOVA Radial	$\left(\sum_{k=1}^n e^{-\sigma(\mathbf{x}_i^k - \mathbf{x}_j^k)^2}\right)^d$
Tangente Hiperbólico	$\tanh(\gamma \mathbf{x}_i^T \mathbf{x}_j)$, $\gamma \geq 0$

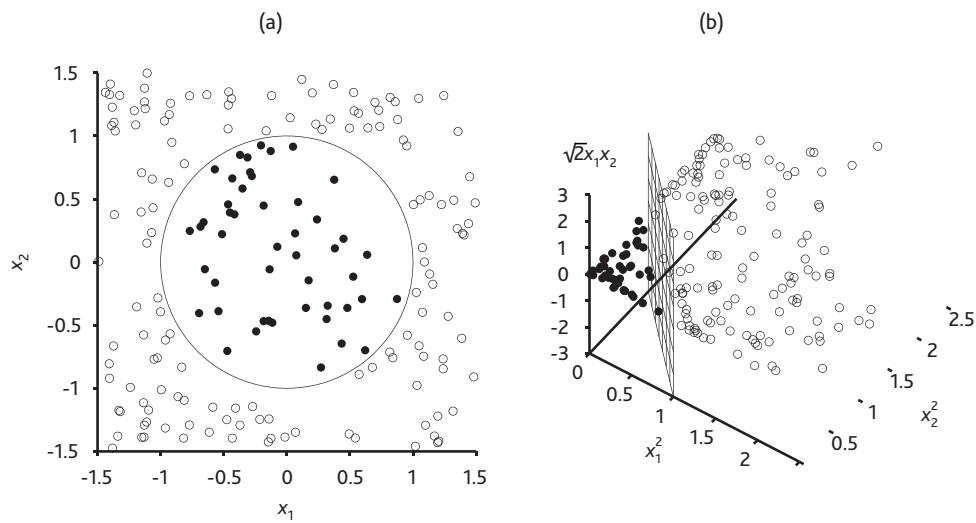


Figura 2.2.: Un ejemplo de SVM en el cual (a) es un conjunto de entrenamiento en el plano (los puntos negros representan los casos) el cual pasa a ser linealmente separable en el espacio tridimensional (b). Modificado de Verplancke et al. (2008).

2.2. Regresión Logística

Sea Y una variable aleatoria tal que

$$Y = \begin{cases} 1, & \text{si la condición está presente} \\ 0, & \text{en otro caso} \end{cases} \quad (2.2)$$

y $\mathbf{x} = (x_1, x_2, \dots, x_p)$ el conjunto de covariables de interés. Se define

$$\pi(\mathbf{x}) = E(Y|x_1, \dots, x_p)$$

como la probabilidad de que una de las observaciones \mathbf{x} pertenezca a uno de los grupos. El modelo de regresión presentado en Hosmer & Lemeshow (1989) es de la forma:

$$\pi(\mathbf{x}) = \frac{\exp\{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p\}}{1 + \exp\{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p\}} \quad (2.3)$$

Aplicando la transformación logit que se define como:

$$\text{logit}(y) = \log(y/(1 - y)) \quad (2.4)$$

en (2.3) se obtiene un modelo lineal en los parámetros. Sea $\hat{\boldsymbol{\beta}}$ el estimador de máxima verosimilitud de $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$. La probabilidad de que una *nueva* observación $\mathbf{x} = (x_1^*, x_2^*, \dots, x_p^*)$ pertenezca a uno de los grupos está dada por

$$\hat{\pi}(x^*) = \frac{\exp\{\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \dots + \hat{\beta}_p x_p^*\}}{1 + \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \dots + \hat{\beta}_p x_p^*\}} \quad (2.5)$$

de tal forma que esta *nueva* observación \mathbf{x}^* será clasificada en el grupo para el cual (2.5) sea mayor.

En la literatura existen otras alternativas para la estimación de los parámetros y la clasificación de un nuevo individuo cuando se utiliza un modelo de RL. En el primer caso por ejemplo, Houston & Woodruff (1997) habla de la implementación de *factores de Bayes* y Piegorisch & Casella (1996) de *estimadores empíricos*

de Bayes. Para la clasificación de nuevos individuos es posible utilizar una generalización de *lasso*¹ o una *regla de Bayes* (Whittemore 1995). En este trabajo sólo se emplearon estimadores de máxima verosimilitud y los nuevos individuos se clasificaron en el grupo para el cual la probabilidad de pertenencia era mayor.

Como puede observarse, el soporte teórico sobre el que reposan ambas metodologías es distinto aunque la idea es similar: clasificar. Hasta ahora, una ventaja aparente de SVM sobre RL es la utilización de diferentes tipos de discriminantes (kernels). Sin embargo, esto no constituye de ninguna manera una evidencia de que SVM sea superior a LR.

¹Least Absolute Shrinkage and Selection Operator.

3. Estrategias de Comparación

En este capítulo se describen las estrategias de simulación que se implementaron en este estudio. En las simulaciones se tuvieron en cuenta diferentes distribuciones para generar simultáneamente los conjuntos de datos de *entrenamiento*¹ y *validación*², donde $g = \{1, 2\}$ representa los grupos mutuamente excluyentes a los que pertenecen los datos.

3.1. Escenarios Univariados

El objetivo de este estudio es evaluar el desempeño de los métodos en escenarios simulados utilizando distribuciones conocidas, con una sola variable predictora, y donde se tiene control de los parámetros. Inicialmente se supone que los dos grupos tienen la misma distribución de probabilidad y se controla el nivel de acercamiento de cada grupo mediante un parámetro d , que define la distancia entre los valores esperados de cada grupo. En general, en todos los escenarios el número de individuos en cada grupo se representa por n_1 y n_2 iguales 20, 50 y 100 y combinaciones de estos. También se considera que las observaciones en cada grupo tienen una distribución de probabilidad diferente y se controla nuevamente el valor esperado de cada grupo, de tal manera que se generan combinaciones de distribuciones y se evalúa si los métodos son sensibles a este tipo de cambios. La tabla 3.1 contiene la información resumida sobre las diferentes distribuciones de probabilidad consideradas, así como valores del parámetro d .

¹Datos de los cuales se conoce su clasificación y son utilizados para construir las funciones discriminantes.

²Datos de los cuales se conoce su clasificación y se utilizan para validar, si las funciones discriminantes clasifican correctamente.

Tabla 3.1.: Distribuciones de probabilidad univariadas consideradas para el estudio.

Distribution	$g = 1$	$g = 2$	d
Poisson	Poisson(1)	Poisson(d)	{3, 5, 8, 10}
Exponencial	Exp(1)	Exp(d)	{3, 5, 8, 10}
Normal	$N(0, 1)$	$N(d, 1)$	{0.5, 1, 2, 2.5}
Lognormal	Lognormal(0, 1)	Lognormal($d, 1$)	$\{\frac{1}{3}, \frac{2}{3}, 1, \frac{3}{2}\}$
Cauchy	Cauchy(0, 1)	Cauchy($d, 1$)	{1, 2, 4, 5}
Cauchy-Normal	Cauchy(0, 1)	$N(d, 1)$	{1, 2, 4, 5}
Normal-Poisson	$N(0, 1)$	Poisson(d)	{1, 2, 4, 5}
Normal-Exponencial	$N(0, 1)$	Exp(d)	$\{2, \frac{1}{2}, \frac{1}{4}, \frac{1}{5}\}$

3.2. Escenarios Normales Multivariantes

Se tienen en cuenta para las simulaciones dos grupos normales multivariados $N_p(\mu_p, \Sigma_{p \times p})$. Los parámetros μ_1, μ_2 , vectores de orden $p \times 1$, son tales que μ_1 permanece constante e igual al vector de ceros, como referencia, y μ_2 se mueve a cuatro distancias $d = 0.5, 1, 2, 2.5$ sobre los ejes coordenados. Σ_1 y Σ_2 son matrices cuadradas de orden p , con $\sigma_i = 1, \forall i \in \{2, \dots, p\}$ y $\sigma_{ij} = \rho_{ij} = 0.1, 0.3, 0.5, 0.7, 0.9$ para $i \neq j, i, j \in \{2, \dots, p\}$.

Normal Bivariada

Considerando el mismo enfoque de las distribuciones normales univariadas donde el propósito principal era evaluar los métodos en contextos teóricos y teniendo en cuenta que en estos escenarios se puede controlar también la correlación entre las covariables, se consideran distribuciones normales bivariadas para los grupos con las siguientes características:

Escenario 1: Dos grupos con número de individuos $n_1 = n_2 = 20, 50, 100$ con $\Sigma_1 = \Sigma_2$ y $\mu_1 = (0, 0)$ permanece constante y μ_2 varía a cuatro distancias d tales que:

1. $\mu_2 = (0, \frac{1}{2})$

2. $\mu_2 = (1, 0)$

3. $\mu_2 = (0, 2)$

4. $\mu_2 = (\frac{5}{2}, 0)$

Escenario 2: Como en el escenario 1, pero con tamaños de muestras para los grupos $n_1 = 20$ $n_2 = 50$, $n_1 = 50$ $n_2 = 100$, y $n_1 = 20$ $n_2 = 100$.

Escenario 3: Igual al escenario 1, pero considerando $\Sigma_1 = 2\Sigma_2$.

Escenario 4: Mismas condiciones que en el escenario 2, suponiendo que $\Sigma_1 = 2\Sigma_2$.

Escenario 5: Se tienen en cuenta la situación del escenario 1, con $\Sigma_1 = 3\Sigma_2$.

Escenario 6: Similar al escenario 2, pero con $\Sigma_1 = 3\Sigma_2$.

Normal Multivariada ($p = 200$)

En este punto se simula una situación donde la cantidad de variables medidas sobre los individuos es mucho mayor que la cantidad de individuos en cada grupo, como ocurre por ejemplo en los estudios genéticos. Se tienen en cuenta los mismos seis escenarios que se construyeron para las distribución normal bivariadas descritas anteriormente, con los cambios esperados debidos al aumento de dimensionalidad en las matrices y vectores.

3.3. Estrategia de Simulación

La estrategia de simulación y comparación involucra los pasos descritos en el siguiente algoritmo.

Algoritmo: Pasos para la Comparación

1. Elija la distribución de probabilidad (univariada ó multivariada).
2. Genere n_g individuos, para formar D , el conjunto de datos de *entrenamiento*.
3. Con D , estime los modelos para RL y SVM.
4. Genere nuevas observaciones como en 1, que serán D^* , el conjunto de datos de *validación*.
5. Evalúe sobre D^* , los modelos estimados en 2. Determine cuántos individuos fueron mal clasificados y calcule la tasa de clasificación errónea (Misclassification Rate [MCR], en inglés).
6. Repita los pasos 4 y 5 $B = 5000$ veces y calcule la MCR promedio.

Los pasos 2-6 fueron programados en R (R Development Core Team 2011) (ver Anexo A). El valor esperado y la varianza de las distribuciones fueron controladas por el parámetro d . Los tamaños muestrales usados fueron (i) $n_1 = n_2 = 20, 50, 100$ y (ii) $n_1 \neq n_2$.

El modelo RL fue ajustado usando la función `glm()` de R y los individuos fueron asignados al grupo g para el cual su probabilidad fuera la más alta. Para el modelo SVM se incluyeron los kernels (i) lineal, (ii), polinomial, (iii) radial y (iv) tangencial, todos ajustados usando la función `tune.svm()` en librería `e1071` (Dimitriadou et al. 2011). Los ajustes o *tuning* de los modelos SVM se realizaron para los parámetros γ , el cual controla la complejidad de la función construida por SVM, y C , que controla la penalidad en la mala clasificación de un punto en el conjunto de entrenamiento (Karatzoglou et al. 2006, pp. 3).

4. Resultados de simulaciones

A continuación se presentan los resultados obtenidos de manera gráfica. De esta forma, se pretende tener de un solo vistazo los resultados y así poder analizarlos de una forma más sistemática y organizada. Sin embargo, si se prefiere conocer el resultado exacto de cada una de las simulaciones, también se presentan los resultados numéricos en forma tabular.

Los resultados aquí presentados se dividen en dos secciones. En la primera se muestran las distribuciones univariadas, donde se supone primero que cada grupo sigue una misma distribución de probabilidad y luego se considera que cada uno sigue una distribución diferente. La segunda sección está dedicada a distribuciones normales p -variantes, con $p = 2, 20, 50, 200$. Sin embargo, sólo se reportaron como resultados $p = 2, 200$ ya que para el resto se observaron comportamientos similares, para mas detalles ver anexo C.

4.1. Univariados

En cada gráfico y tabla se muestra la MCR como función de d para los modelos RL y SVM cuando las observaciones provienen de una distribución Normal, Poisson, Exponencial, Cauchy y Lognormal. Los tamaños de las muestras en el panel superior de cada gráfico, denotado por (i) , son iguales con (a) 20, (b) 50 y (c) 100 individuos por grupo. En el panel inferior, denotado por (ii) , los tamaños de muestra utilizados fueron (a) $n_1 = 20, n_2 = 50$, (b) $n_1 = 50, n_2 = 100$, (c) $n_1 = 20, n_2 = 100$. Ver tabla 3.1 para más detalles.

4.1.1. Distribución Normal

De acuerdo con el panel superior de la figura 4.1, se puede argumentar que el kernel polinomial tiene un desempeño pobre en comparación con los demás. A medida que aumenta el tamaño muestral, los métodos RL y SVM con kernel lineal, radial y tangencial son prácticamente idénticos, lo que sugiere que el aumento en el tamaño de los grupos no altera comportamiento de los métodos. Como era de esperarse, la tasa de clasificación errónea disminuye a medida que d aumenta. Cuando la cantidad de individuos en cada grupo es diferente, se observa una leve alteración en el kernel tangencial, en tanto el kernel polinomial mejora su desempeño en relación con los demás. Bajo estas condiciones las menores tasas son producidas cuando la diferencia entre los individuos en cada grupo es la mayor (véase panel (c), por ejemplo).

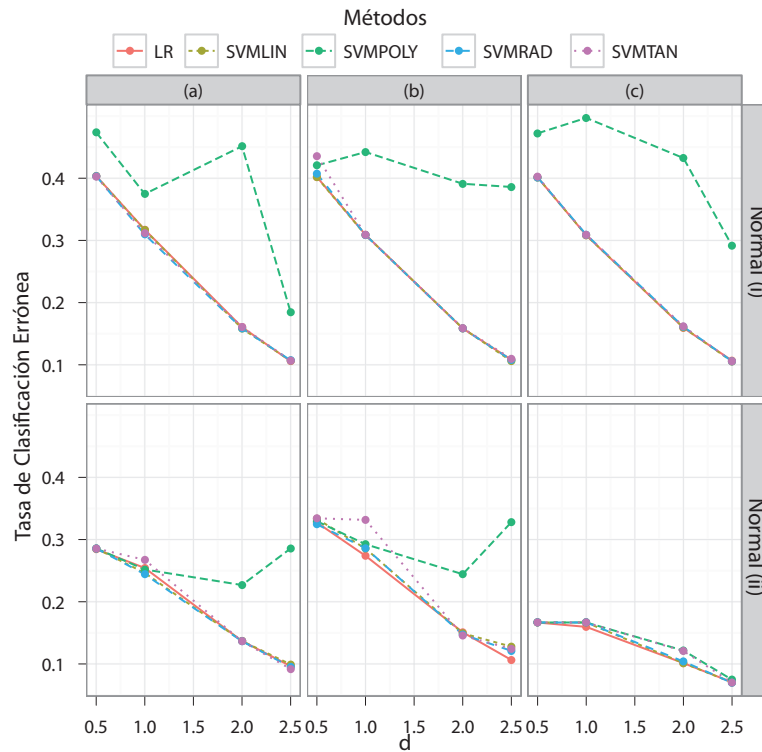


Figura 4.1.: Distribución Normal

Tabla 4.1.: Resultados distribución Normal.

$n_1 = n_2 = 20 ((a), i)$					
DIS / MET	RL	SVMLIN	SVMPOLY	SVMRAD	SVMTAN
Distancia 1	0.40	0.40	0.47	0.40	0.40
Distancia 2	0.32	0.32	0.37	0.31	0.31
Distancia 3	0.16	0.16	0.45	0.16	0.16
Distancia 4	0.11	0.11	0.18	0.11	0.11
$n_1 = n_2 = 50 ((b), i)$					
Distancia 1	0.40	0.40	0.42	0.41	0.44
Distancia 2	0.31	0.31	0.44	0.31	0.31
Distancia 3	0.16	0.16	0.39	0.16	0.16
Distancia 4	0.11	0.11	0.39	0.11	0.11
$n_1 = n_2 = 100 ((c), i)$					
Distancia 1	0.40	0.40	0.47	0.40	0.40
Distancia 2	0.31	0.31	0.50	0.31	0.31
Distancia 3	0.16	0.16	0.43	0.16	0.16
Distancia 4	0.11	0.11	0.29	0.11	0.11
$n_1 = 20 \text{ y } n_2 = 50 ((a), ii)$					
Distancia 1	0.28	0.29	0.29	0.29	0.29
Distancia 2	0.25	0.25	0.25	0.24	0.27
Distancia 3	0.14	0.14	0.23	0.14	0.14
Distancia 4	0.10	0.10	0.29	0.10	0.09
$n_1 = 50, n_2 = 100 ((b), ii)$					
Distancia 1	0.33	0.33	0.33	0.32	0.33
Distancia 2	0.27	0.29	0.29	0.29	0.33
Distancia 3	0.15	0.15	0.24	0.15	0.15
Distancia 4	0.11	0.13	0.33	0.12	0.12
$n_1 = 20, n_2 = 100 ((c), ii)$					
Distancia 1	0.17	0.17	0.17	0.17	0.17
Distancia 2	0.16	0.17	0.17	0.17	0.17
Distancia 3	0.10	0.10	0.12	0.10	0.12
Distancia 4	0.07	0.07	0.07	0.07	0.07

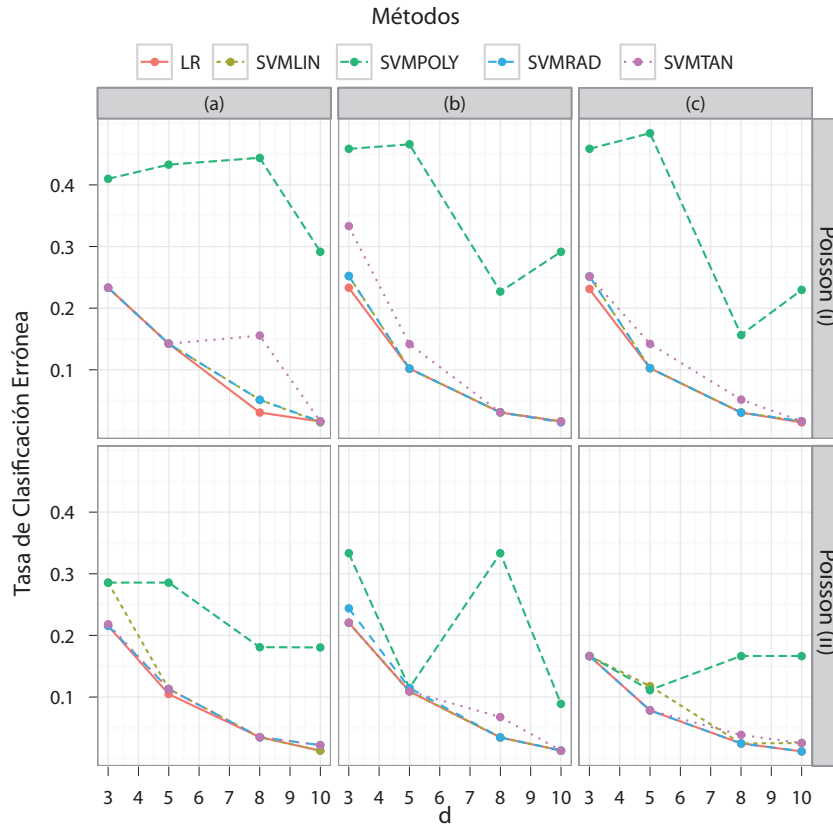


Figura 4.2.: Distribución Poisson

4.1.2. Distribución Poisson

En la figura 4.2 se observa un desempeño pobre del kernel polinomial respecto de los demás, mientras el kernel lineal, radial y la regresión logística disminuyen sus tasas de mala clasificación a medida que d aumenta con un desempeño similar entre ellos. Por otra parte, la cantidad de individuos no parece alterar el comportamiento de los métodos cuando $n_1 = n_2$, pero si se observa un leve mejoría cuando $n_1 \neq n_2$. Finalmente se puede afirmar que el kernel tangencial no se desempeña tan bien como el radial, lineal y RL.

Tabla 4.2.: Resultados distribución Poisson.

$n_1 = n_2 = 20 ((a), i)$					
DIS / MET	RL	SVMLIN	SVMPOLY	SVMRAD	SVMTAN
Distancia 1	0.23	0.23	0.41	0.23	0.23
Distancia 2	0.14	0.14	0.43	0.14	0.14
Distancia 3	0.03	0.05	0.44	0.05	0.16
Distancia 4	0.02	0.01	0.29	0.02	0.02
$n_1 = n_2 = 50 ((b), i)$					
Distancia 1	0.23	0.25	0.46	0.25	0.33
Distancia 2	0.10	0.10	0.47	0.10	0.14
Distancia 3	0.03	0.03	0.23	0.03	0.03
Distancia 4	0.02	0.02	0.29	0.01	0.02
$n_1 = n_2 = 100 ((c), i)$					
Distancia 1	0.23	0.25	0.46	0.25	0.25
Distancia 2	0.10	0.10	0.48	0.10	0.14
Distancia 3	0.03	0.03	0.16	0.03	0.05
Distancia 4	0.01	0.02	0.23	0.02	0.02
$n_1 = 20, n_2 = 50 ((a), ii)$					
Distancia 1	0.22	0.29	0.29	0.22	0.22
Distancia 2	0.10	0.11	0.29	0.11	0.11
Distancia 3	0.04	0.04	0.18	0.04	0.04
Distancia 4	0.01	0.01	0.18	0.02	0.02
$n_1 = 50, n_2 = 100 ((b), ii)$					
Distancia 1	0.22	0.22	0.33	0.24	0.22
Distancia 2	0.11	0.11	0.11	0.11	0.11
Distancia 3	0.03	0.03	0.33	0.03	0.07
Distancia 4	0.01	0.01	0.09	0.01	0.01
$n_1 = 20, n_2 = 100 ((c), ii)$					
Distancia 1	0.17	0.17	0.17	0.17	0.17
Distancia 2	0.08	0.12	0.11	0.08	0.08
Distancia 3	0.02	0.02	0.17	0.02	0.04
Distancia 4	0.01	0.03	0.17	0.01	0.03

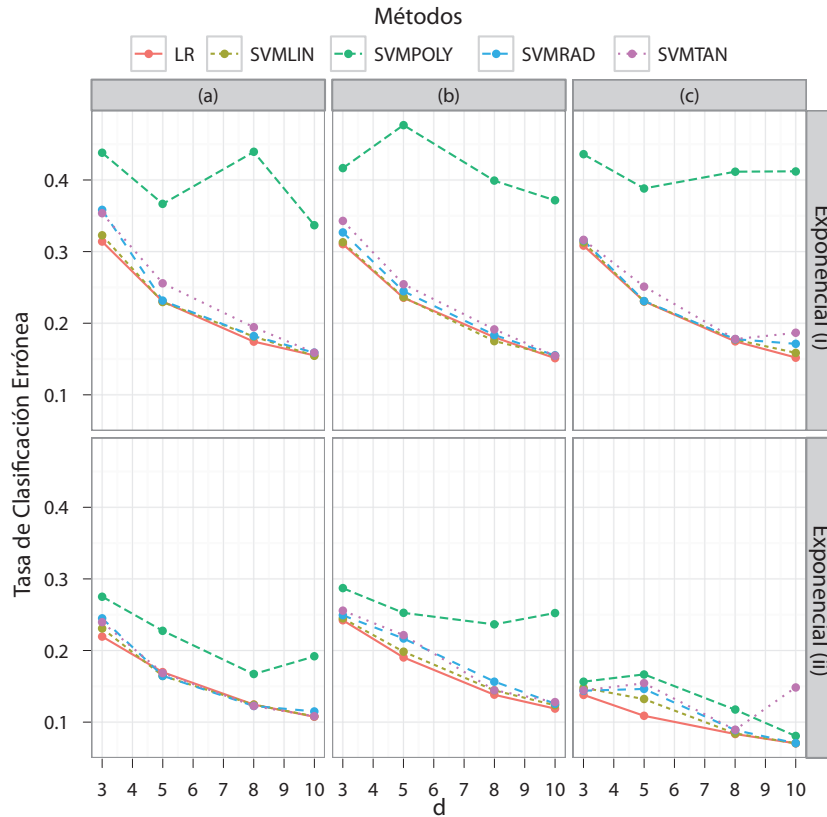


Figura 4.3.: Distribución Exponencial

4.1.3. Distribución Exponencial

De acuerdo con los resultados de la figura 4.3 se puede afirmar que cuando los tamaños son iguales, el SVM compite bien con RL excepto por el kernel polinomial. Se evidencia un comportamiento similar en los métodos sin importar el aumento en la cantidad de individuos en los grupos (especialmente cuando $n_1 = n_2$), obteniéndose tasas no inferiores a 0.14. Cuando los tamaños muestrales de cada grupo difieren, la MCRs en la mayoría de los métodos disminuye, y se alcanzan MCRs inferiores a 0.14 cuando $n_1 = 20$ y $n_2 = 100$. Sólo en este caso se observa que LR aparentemente es mejor que SVM. De nuevo el kernel polinomial no se recomienda, debido a su pobre desempeño en términos de la MCR.

Tabla 4.3.: Resultados distribución Exponencial.

$n_1 = n_2 = 20 ((a), i)$					
DIS / MET	RL	SVMLIN	SVMPOLY	SVMRAD	SVMTAN
Distancia 1	0.31	0.32	0.44	0.36	0.35
Distancia 2	0.23	0.23	0.37	0.23	0.26
Distancia 3	0.17	0.18	0.44	0.18	0.19
Distancia 4	0.16	0.15	0.34	0.16	0.16
$n_1 = n_2 = 50 ((b), i)$					
Distancia 1	0.31	0.31	0.42	0.33	0.34
Distancia 2	0.24	0.24	0.48	0.24	0.25
Distancia 3	0.18	0.18	0.40	0.18	0.19
Distancia 4	0.15	0.15	0.37	0.15	0.15
$n_1 = n_2 = 100 ((c), i)$					
Distancia 1	0.31	0.31	0.44	0.32	0.32
Distancia 2	0.23	0.23	0.39	0.23	0.25
Distancia 3	0.17	0.18	0.41	0.18	0.18
Distancia 4	0.15	0.16	0.41	0.17	0.19
$n_1 = 20, n_2 = 50 ((a), ii)$					
Distancia 1	0.22	0.23	0.28	0.24	0.24
Distancia 2	0.17	0.16	0.23	0.16	0.17
Distancia 3	0.12	0.12	0.17	0.12	0.12
Distancia 4	0.11	0.11	0.19	0.11	0.11
$n_1 = 50, n_2 = 100 ((b), ii)$					
Distancia 1	0.24	0.24	0.29	0.25	0.26
Distancia 2	0.19	0.20	0.25	0.22	0.22
Distancia 3	0.14	0.14	0.24	0.16	0.14
Distancia 4	0.12	0.12	0.25	0.13	0.13
$n_1 = 20, n_2 = 100 ((c), ii)$					
Distancia 1	0.14	0.15	0.16	0.14	0.14
Distancia 2	0.11	0.13	0.17	0.15	0.15
Distancia 3	0.08	0.08	0.12	0.09	0.09
Distancia 4	0.07	0.07	0.08	0.07	0.15

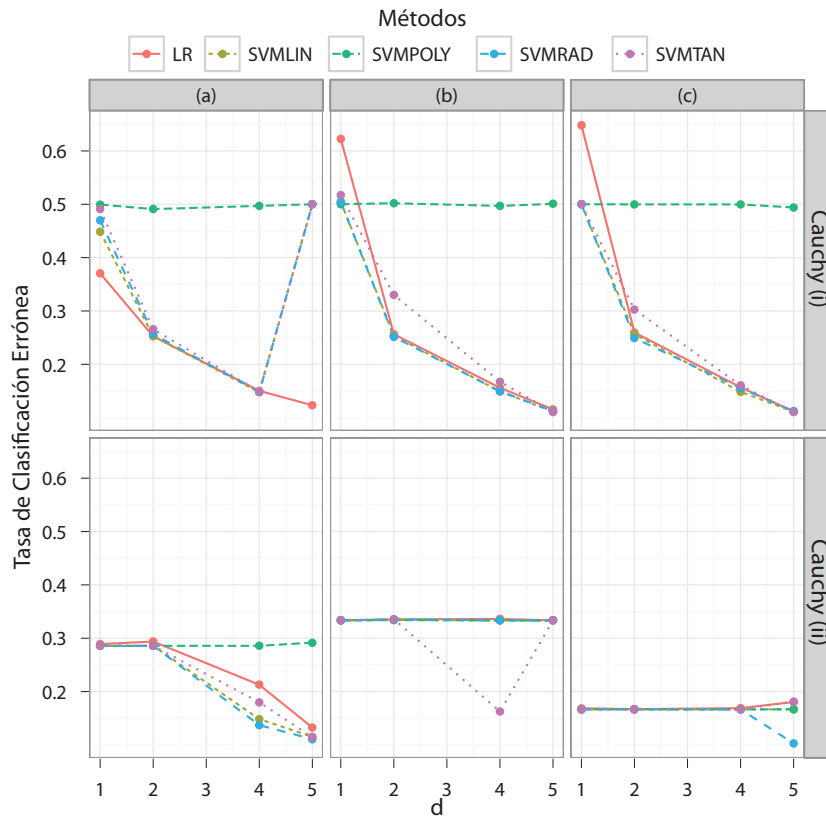


Figura 4.4.: Distribución Cauchy

4.1.4. Distribución Cauchy

Nuevamente en la figura 4.4 el kernel polinomial exhibe un mal desempeño. Sin embargo, se observa que en todos los casos, las SVM igualan por poco o mejoran la MCR dada por la RL. Con base en estos resultados se puede afirmar que bajo una distribución Cauchy es recomendable usar SVM como método de clasificación sobre todo cuando la cantidad de individuos en cada grupo es diferente o los grupos estén muy mezclados (d pequeño). Así mismo, en el panel superior de la figura 4.4 se observa que las MCR obtenidas por los métodos evaluados parecen no verse afectadas por el incremento en el número de individuos en cada grupo.

Tabla 4.4.: Resultados distribución Cauchy.

$n_1 = n_2 = 20 ((a), i)$					
DIS / MET	RL	SVMLIN	SVMPOLY	SVMRAD	SVMTAN
Distancia 1	0.37	0.45	0.50	0.47	0.49
Distancia 2	0.25	0.25	0.49	0.26	0.27
Distancia 3	0.15	0.15	0.50	0.15	0.15
Distancia 4	0.12	0.50	0.50	0.50	0.50
$n_1 = n_2 = 50 ((b), i)$					
Distancia 1	0.62	0.50	0.50	0.50	0.52
Distancia 2	0.26	0.25	0.50	0.25	0.33
Distancia 3	0.16	0.15	0.50	0.15	0.17
Distancia 4	0.12	0.11	0.50	0.11	0.11
$n_1 = n_2 = 100 ((c), i)$					
Distancia 1	0.65	0.50	0.50	0.50	0.50
Distancia 2	0.26	0.26	0.50	0.25	0.30
Distancia 3	0.16	0.15	0.50	0.16	0.16
Distancia 4	0.11	0.11	0.49	0.11	0.11
$n_1 = 20, n_2 = 50 ((a), ii)$					
Distancia 1	0.29	0.29	0.29	0.29	0.29
Distancia 2	0.29	0.29	0.29	0.29	0.29
Distancia 3	0.21	0.15	0.29	0.14	0.18
Distancia 4	0.13	0.12	0.29	0.11	0.11
$n_1 = 50, n_2 = 100 ((b), ii)$					
Distancia 1	0.33	0.33	0.33	0.33	0.33
Distancia 2	0.34	0.33	0.33	0.34	0.34
Distancia 3	0.34	0.33	0.33	0.33	0.16
Distancia 4	0.33	0.33	0.33	0.33	0.33
$n_1 = 20, n_2 = 100 ((c), ii)$					
Distancia 1	0.17	0.17	0.17	0.17	0.17
Distancia 2	0.17	0.17	0.17	0.17	0.17
Distancia 3	0.17	0.17	0.17	0.17	0.17
Distancia 4	0.18	0.17	0.17	0.10	0.18

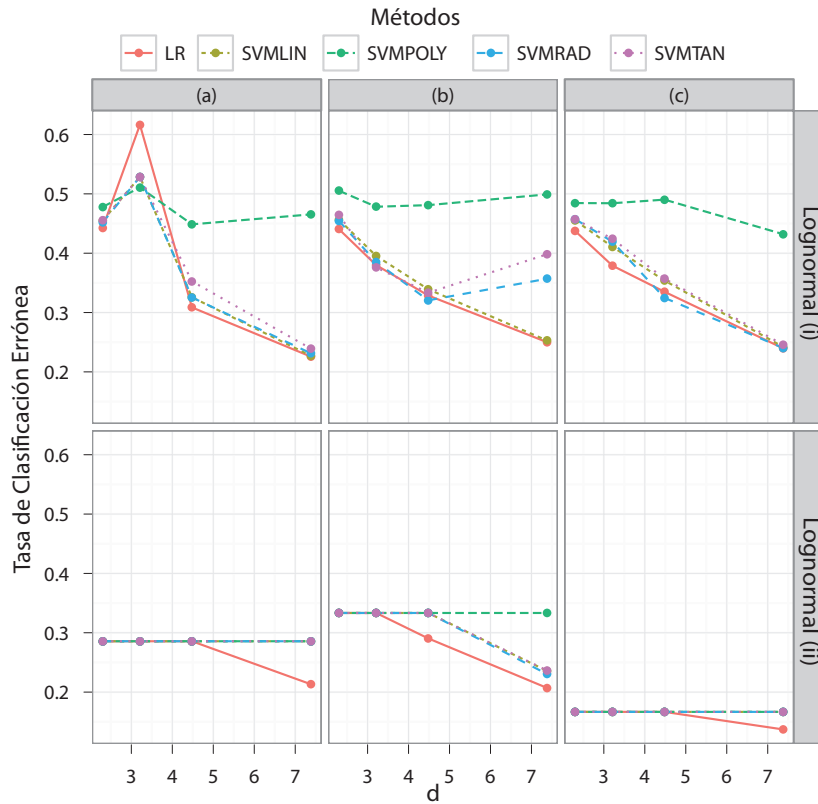


Figura 4.5.: Distribución Lognormal

4.1.5. Distribución Lognormal

En la figura 4.5 se observa que cuando la cantidad de individuos en cada grupo es la menor ($n_1 = n_2 = 20$) y la distancia es pequeña, SVM es mejor alternativa que RL. En los demás casos RL supera o iguala el desempeño de SVM.

Tabla 4.5.: Resultados distribución Lognormal.

$n_1 = n_2 = 20 ((a), i)$					
DIS / MET	RL	SVMLIN	SVMPOLY	SVMRAD	SVMTAN
Distancia 1	0.44	0.46	0.48	0.45	0.46
Distancia 2	0.62	0.53	0.51	0.53	0.53
Distancia 3	0.31	0.33	0.45	0.33	0.35
Distancia 4	0.23	0.23	0.47	0.23	0.24
$n_1 = n_2 = 50 ((b), i)$					
Distancia 1	0.44	0.46	0.51	0.45	0.46
Distancia 2	0.38	0.40	0.48	0.39	0.38
Distancia 3	0.33	0.34	0.48	0.32	0.33
Distancia 4	0.25	0.25	0.50	0.36	0.40
$n_1 = n_2 = 100 ((c), i)$					
Distancia 1	0.44	0.46	0.48	0.46	0.46
Distancia 2	0.38	0.41	0.48	0.42	0.42
Distancia 3	0.33	0.35	0.49	0.32	0.36
Distancia 4	0.24	0.24	0.43	0.24	0.25
$n_1 = 20, n_2 = 50 ((a), ii)$					
Distancia 1	0.29	0.29	0.29	0.29	0.29
Distancia 2	0.29	0.29	0.29	0.29	0.29
Distancia 3	0.29	0.29	0.29	0.29	0.29
Distancia 4	0.21	0.29	0.29	0.29	0.29
$n_1 = 50, n_2 = 100 ((b), ii)$					
Distancia 1	0.33	0.33	0.33	0.33	0.33
Distancia 2	0.33	0.33	0.33	0.33	0.33
Distancia 3	0.29	0.33	0.33	0.33	0.33
Distancia 4	0.21	0.24	0.33	0.23	0.24
$n_1 = 20, n_2 = 100 ((c), ii)$					
Distancia 1	0.17	0.17	0.17	0.17	0.17
Distancia 2	0.17	0.17	0.17	0.17	0.17
Distancia 3	0.17	0.17	0.17	0.17	0.17
Distancia 4	0.14	0.17	0.17	0.17	0.17

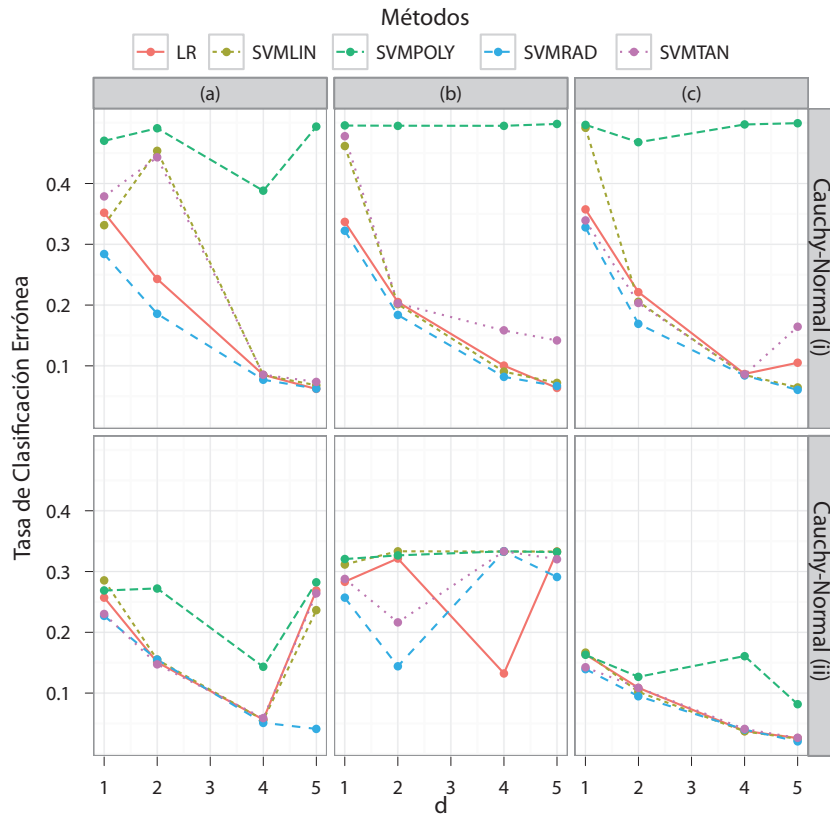


Figura 4.6.: Distribución Cauchy-Normal

4.2. Combinación de Distribuciones

4.2.1. Distribución Cauchy-Normal

En la figura 4.6 continúa el mal desempeño del kernel polinomial cuando la cantidad de individuos en cada grupo es igual (panel superior). Sin embargo, se nota un mejor desempeño de este kernel cuando se tienen menos individuos en el grupo con distribución Cauchy (panel inferior). Es evidente que el comportamiento de las SVM mejora frente a la RL, pues en la mayoría de los casos el kernel radial mejora o iguala los resultados obtenidos con LR. Nuevamente no parece haber diferencia en los métodos al aumentar simultáneamente n_1 y n_2 .

Tabla 4.6.: Resultados distribución Cauchy-Normal.

$n_1 = n_2 = 20 ((a), i)$					
DIS / MET	RL	SVMLIN	SVMPOLY	SVMRAD	SVMTAN
Distancia 1	0.35	0.33	0.47	0.28	0.38
Distancia 2	0.24	0.45	0.49	0.19	0.44
Distancia 3	0.09	0.08	0.39	0.08	0.09
Distancia 4	0.06	0.07	0.49	0.06	0.07
$n_1 = n_2 = 50 ((b), i)$					
Distancia 1	0.34	0.46	0.50	0.32	0.48
Distancia 2	0.20	0.20	0.50	0.18	0.20
Distancia 3	0.10	0.09	0.49	0.08	0.16
Distancia 4	0.06	0.07	0.50	0.07	0.14
$n_1 = n_2 = 100 ((c), i)$					
Distancia 1	0.36	0.49	0.50	0.33	0.34
Distancia 2	0.22	0.21	0.47	0.17	0.20
Distancia 3	0.09	0.09	0.50	0.08	0.09
Distancia 4	0.11	0.06	0.50	0.06	0.16
$n_1 = 20, n_2 = 50 ((a), ii)$					
Distancia 1	0.26	0.29	0.27	0.23	0.23
Distancia 2	0.15	0.15	0.27	0.16	0.15
Distancia 3	0.06	0.06	0.14	0.05	0.06
Distancia 4	0.27	0.24	0.28	0.04	0.26
$n_1 = 50, n_2 = 100 ((b), ii)$					
Distancia 1	0.28	0.31	0.32	0.26	0.29
Distancia 2	0.32	0.33	0.33	0.14	0.22
Distancia 3	0.13	0.33	0.33	0.33	0.33
Distancia 4	0.33	0.33	0.33	0.29	0.32
$n_1 = 20, n_2 = 100 ((c), ii)$					
Distancia 1	0.16	0.17	0.16	0.14	0.14
Distancia 2	0.11	0.10	0.13	0.09	0.11
Distancia 3	0.04	0.04	0.16	0.04	0.04
Distancia 4	0.03	0.02	0.08	0.02	0.03

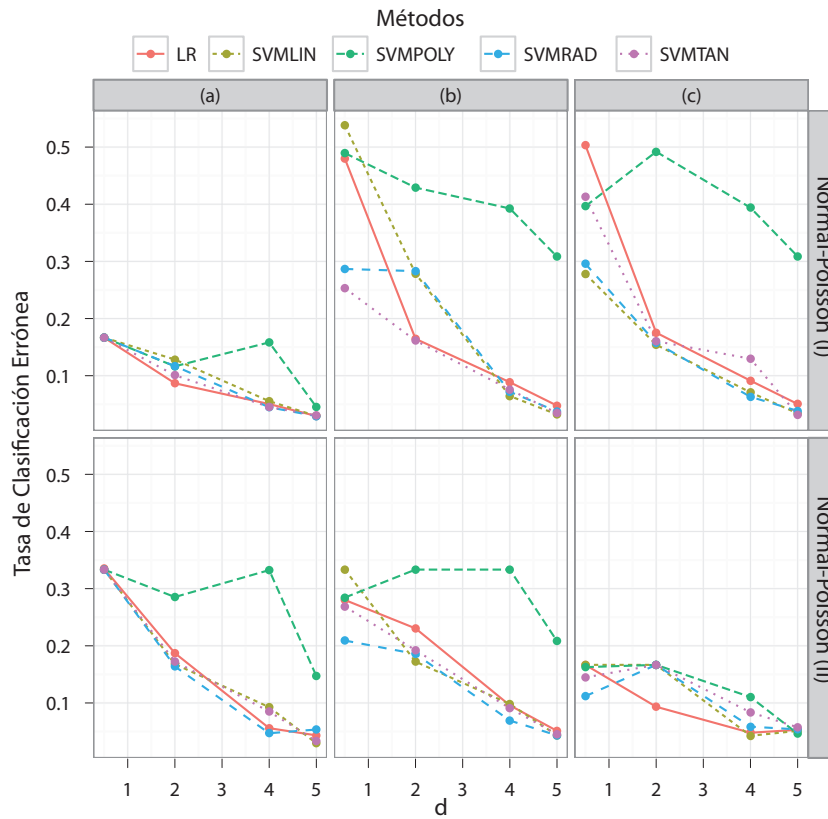


Figura 4.7.: Distribución Normal-Poisson

4.2.2. Distribución Normal-Poisson

En este escenario (figura 4.7) se observa una mejoría del desempeño de las SVM frente a la RL, particularmente cuando el valor esperado de ambos grupos es similar. A medida que aumenta el tamaño muestral, las SVM con kernels lineal y radial superan a LR cuando la cantidad de individuos en cada grupo es igual. En general, para este escenario se tienen mejores resultados en todos los métodos SVM cuando $n_1 = 20$ y $n_2 = 100$.

Tabla 4.7.: Resultados distribución Normal-Poisson.

$n_1 = n_2 = 20 ((a), i)$					
DIS / MET	RL	SVMLIN	SVMPOLY	SVMRAD	SVMTAN
Distancia 1	0.17	0.17	0.17	0.17	0.17
Distancia 2	0.09	0.13	0.12	0.12	0.10
Distancia 3	0.05	0.06	0.16	0.04	0.05
Distancia 4	0.03	0.03	0.05	0.03	0.03
$n_1 = n_2 = 50 ((b), i)$					
Distancia 1	0.48	0.54	0.49	0.29	0.25
Distancia 2	0.16	0.28	0.43	0.28	0.16
Distancia 3	0.09	0.06	0.39	0.07	0.08
Distancia 4	0.05	0.03	0.31	0.04	0.03
$n_1 = n_2 = 100 ((c), i)$					
Distancia 1	0.50	0.28	0.40	0.30	0.41
Distancia 2	0.18	0.15	0.49	0.16	0.16
Distancia 3	0.09	0.07	0.39	0.06	0.13
Distancia 4	0.05	0.03	0.31	0.04	0.03
$n_1 = 20, n_2 = 50 ((a), ii)$					
Distancia 1	0.34	0.33	0.33	0.33	0.33
Distancia 2	0.19	0.17	0.29	0.16	0.17
Distancia 3	0.06	0.09	0.33	0.05	0.08
Distancia 4	0.04	0.03	0.15	0.05	0.03
$n_1 = 50, n_2 = 100 ((b), ii)$					
Distancia 1	0.28	0.33	0.28	0.21	0.27
Distancia 2	0.23	0.17	0.33	0.19	0.19
Distancia 3	0.10	0.10	0.33	0.07	0.09
Distancia 4	0.05	0.04	0.21	0.04	0.05
$n_1 = 20, n_2 = 100 ((c), ii)$					
Distancia 1	0.17	0.17	0.16	0.11	0.14
Distancia 2	0.09	0.17	0.17	0.17	0.17
Distancia 3	0.05	0.04	0.11	0.06	0.08
Distancia 4	0.05	0.05	0.05	0.05	0.06

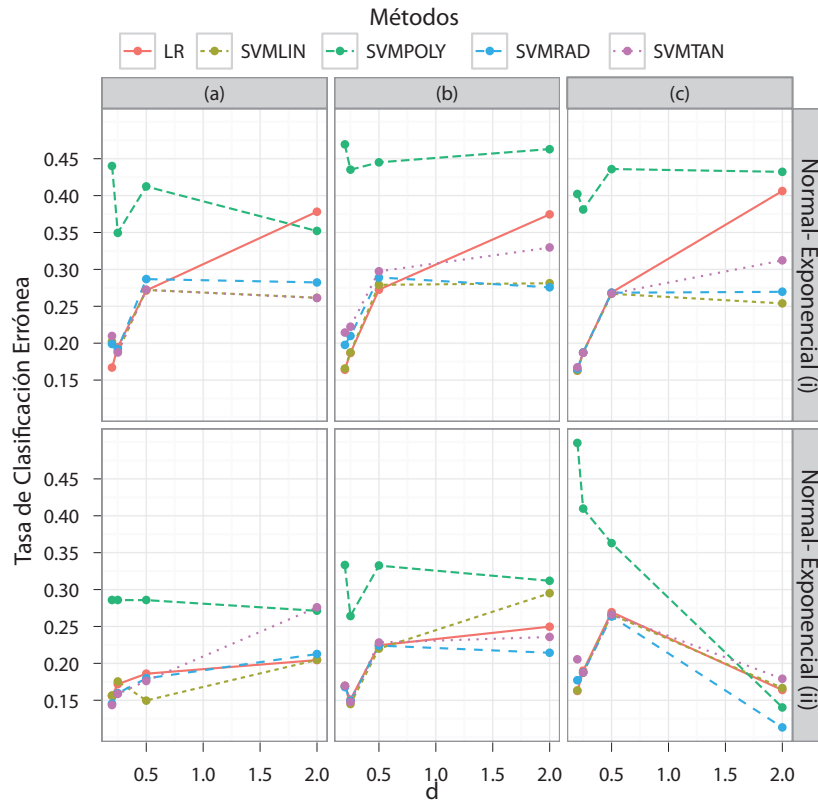


Figura 4.8.: Distribución Normal-Exponencial

4.2.3. Distribución Normal-Exponencial

Es de resaltar que por la naturaleza de las distribuciones, grandes valores de d implican que los grupos se encuentran más cerca. De acuerdo a los resultados presentados en la figura 4.8, a excepción del kernel polinomial, las SVM representan una alternativa considerable para reemplazar a RL cuando los grupos están más cerca.

Tabla 4.8.: Resultados distribución Normal-Exponencial.

$n_1 = n_2 = 20 ((a), i)$					
DIS / MET	RL	SVMLIN	SVMPOLY	SVMRAD	SVMTAN
Distancia 1	0.38	0.26	0.35	0.28	0.26
Distancia 2	0.27	0.27	0.41	0.29	0.27
Distancia 3	0.19	0.19	0.35	0.19	0.19
Distancia 4	0.17	0.20	0.44	0.20	0.21
$n_1 = n_2 = 50 ((b), i)$					
Distancia 1	0.37	0.28	0.46	0.28	0.33
Distancia 2	0.27	0.28	0.45	0.29	0.30
Distancia 3	0.19	0.19	0.44	0.21	0.22
Distancia 4	0.16	0.17	0.47	0.20	0.21
$n_1 = n_2 = 100 ((c), i)$					
Distancia 1	0.41	0.25	0.43	0.27	0.31
Distancia 2	0.27	0.27	0.44	0.27	0.27
Distancia 3	0.19	0.19	0.38	0.19	0.19
Distancia 4	0.16	0.16	0.40	0.17	0.17
$n_1 = 20, n_2 = 50 ((a), ii)$					
Distancia 1	0.20	0.20	0.27	0.21	0.28
Distancia 2	0.19	0.15	0.29	0.18	0.18
Distancia 3	0.17	0.18	0.29	0.16	0.16
Distancia 4	0.16	0.16	0.29	0.15	0.14
$n_1 = 50, n_2 = 100 ((b), ii)$					
Distancia 1	0.25	0.30	0.31	0.21	0.24
Distancia 2	0.22	0.22	0.33	0.22	0.23
Distancia 3	0.15	0.14	0.26	0.15	0.15
Distancia 4	0.17	0.17	0.33	0.17	0.17
$n_1 = 20, n_2 = 100 ((c), ii)$					
Distancia 1	0.16	0.17	0.14	0.11	0.18
Distancia 2	0.27	0.27	0.36	0.26	0.27
Distancia 3	0.19	0.19	0.41	0.19	0.19
Distancia 4	0.16	0.16	0.50	0.18	0.21

4.3. Distribuciones Multivariadas

En cada gráfico y tabla se muestra la MCR como función de ρ para los modelos RL y SVM cuando los individuos provienen de una distribución p -variada. Para $p = 2$, se consideraron los vectores de medias (a) (0,0), (b) (1,0), (c) (1, 1.5) y (d) (2.5, 0). Las filas en los gráficos corresponden a combinaciones de tamaños de muestra de la forma (n_1, n_2) , por lo que (20, 50) corresponde $n_1 = 20$ y $n_2 = 50$.

4.3.1. Distribución Normal Bivariada

En la figura 4.9, la MCR tiende a disminuir a medida que la diferencia entre los vectores de medias y la correlación entre variables aumenta. Sin embargo, las MCRs son similares y con pocas variaciones como función de ρ para todos los métodos cuando la cantidad de individuos en cada grupo es diferente y los vectores de media están más cerca. Bajo normalidad bivariada, el desempeño de los kernels lineal, radial y tangencial es bueno, con resultados muy similares a los obtenidos con RL en la mayoría de los casos, mientras que el kernel polinomial sigue mostrando un pobre desempeño.

En el caso en que se considera que las matrices de varianzas y covarianzas en cada grupo son diferentes (figuras 4.10 y 4.11) se evidencia una mejoría sustancial de las SVM con kernel radial sobre RL, especialmente cuando la distancia considerada entre los grupos es la menor y la cantidad de individuos es la misma.

4.3.2. Distribución Normal Multivariada con $p = 200$

Cuando se supone que la cantidad de covariables o variables explicativas en los conjuntos de entrenamiento y validación son mayores que la cantidad de individuos en cada grupo, (figuras 4.12 a 4.14) se evidencia una leve tendencia a aumentar del MCR producido por los todos métodos cuando ρ y d es mayor. No obstante, se obtienen tasas inferiores a las que se tenían con la normal bivariada. En cuanto al desempeño de los métodos entre si, este no cambia en relación a los obtenidos para $p = 2$, lo que permite concluir que p parece no afectar considerablemente el desempeño de los métodos SVM para clasificar. En general se observan resultados

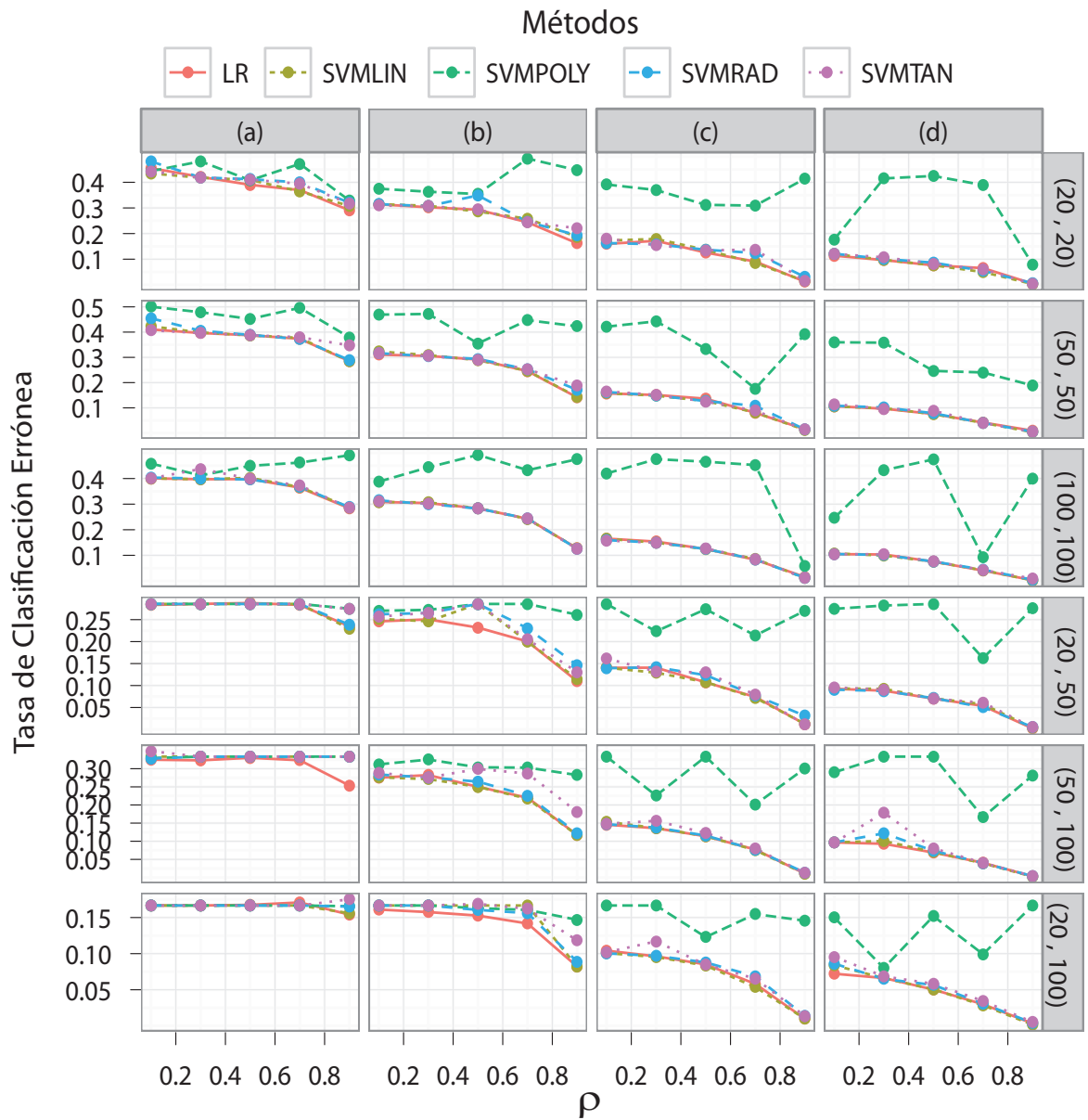


Figura 4.9.: Distribución Normal Bivariada $\Sigma_1 = \Sigma_2$

muy similares entre SVM (kernels lineal y radial) frente a RL. Este resultado es consistente con Shou et al. (2009).

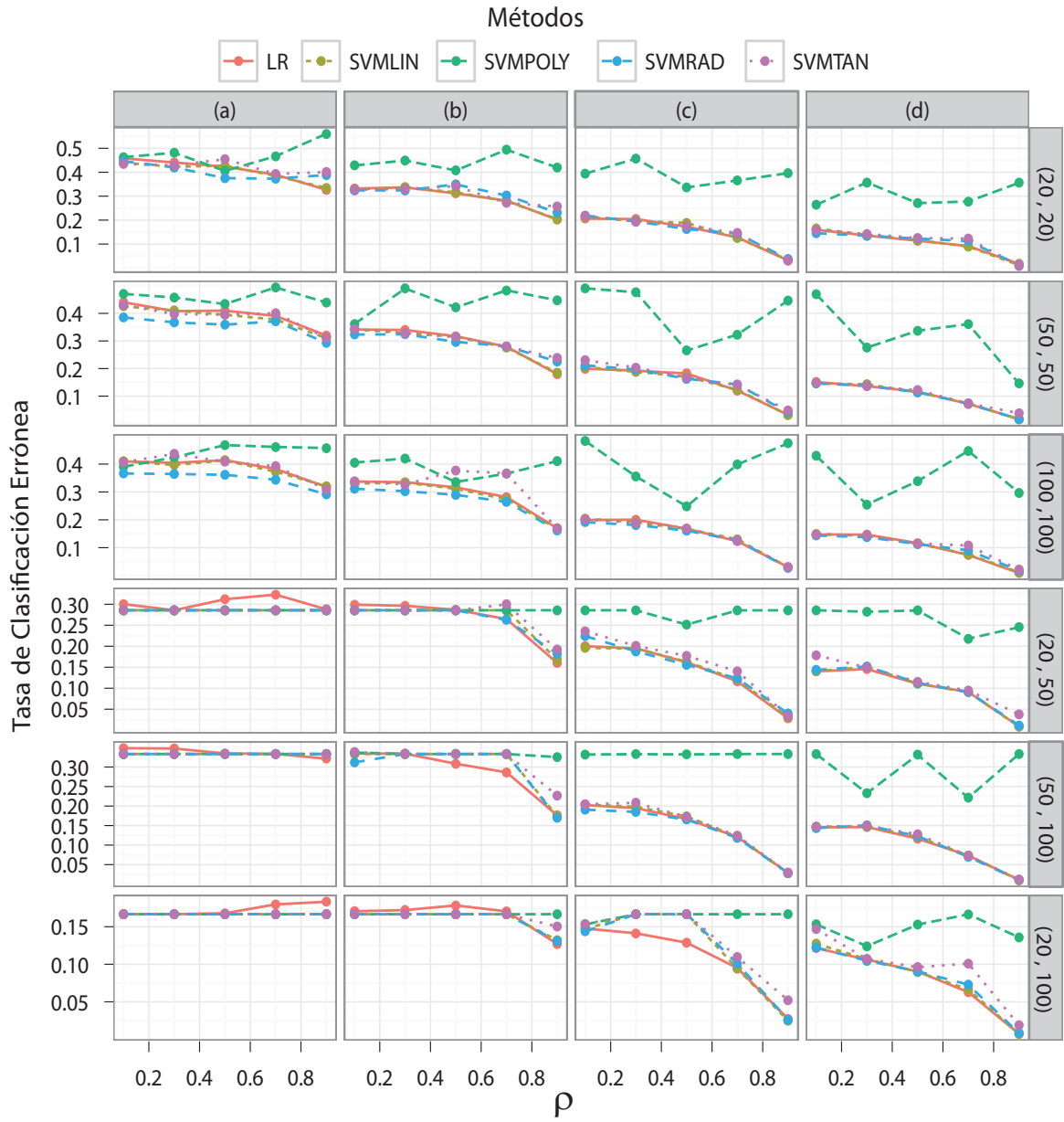


Figura 4.10.: Distribución Normal Bivariada $\Sigma_1 = 2\Sigma_2$ diferente

Tabla 4.9.: Resultados distribución Normal Bivariada $\Sigma_1 = \Sigma_2$.

MÉTODO/ ρ		0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
Distancia 1	$n_1 = n_2 = 20$						$n_1 = 20, n_2 = 50$				
	RL	0.455	0.421	0.390	0.370	0.291	0.283	0.286	0.288	0.283	0.235
	SVMLIN	0.435	0.418	0.410	0.365	0.308	0.286	0.286	0.286	0.286	0.229
	SVMPOLY	0.444	0.482	0.408	0.472	0.329	0.286	0.286	0.286	0.286	0.275
	SVMRAD	0.481	0.419	0.412	0.400	0.320	0.286	0.286	0.286	0.286	0.238
Distancia 2	SVMTAN	0.442	0.418	0.411	0.394	0.317	0.286	0.286	0.286	0.286	0.274
	RL	0.314	0.303	0.293	0.244	0.162	0.246	0.251	0.232	0.200	0.110
	SVMLIN	0.315	0.309	0.286	0.258	0.187	0.252	0.246	0.286	0.200	0.115
	SVMPOLY	0.375	0.363	0.354	0.493	0.447	0.270	0.272	0.286	0.286	0.261
	SVMRAD	0.315	0.307	0.349	0.246	0.193	0.262	0.265	0.286	0.230	0.146
Distancia 3	SVMTAN	0.309	0.307	0.294	0.246	0.221	0.257	0.266	0.286	0.205	0.131
	RL	0.159	0.172	0.126	0.091	0.013	0.140	0.141	0.107	0.074	0.012
	SVMLIN	0.173	0.179	0.137	0.085	0.015	0.141	0.129	0.108	0.072	0.013
	SVMPOLY	0.391	0.370	0.312	0.309	0.414	0.286	0.223	0.274	0.214	0.270
	SVMRAD	0.163	0.158	0.137	0.124	0.031	0.139	0.141	0.124	0.077	0.031
Distancia 4	SVMTAN	0.180	0.155	0.134	0.137	0.016	0.162	0.132	0.130	0.079	0.011
	RL	0.114	0.096	0.076	0.065	0.005	0.093	0.088	0.070	0.054	0.002
	SVMLIN	0.120	0.098	0.076	0.050	0.003	0.093	0.092	0.071	0.058	0.004
	SVMPOLY	0.177	0.415	0.425	0.389	0.080	0.275	0.282	0.286	0.162	0.276
	SVMRAD	0.121	0.101	0.086	0.055	0.006	0.091	0.088	0.071	0.051	0.005
SVMTAN	0.120	0.108	0.082	0.061	0.003	0.096	0.091	0.070	0.061	0.005	
Distancia 1	$n_1 = n_2 = 50$						$n_1 = 50, n_2 = 100$				
	RL	0.410	0.396	0.388	0.374	0.285	0.325	0.323	0.330	0.323	0.254
	SVMLIN	0.423	0.398	0.387	0.377	0.284	0.333	0.333	0.333	0.333	0.333
	SVMPOLY	0.500	0.479	0.452	0.496	0.379	0.329	0.333	0.333	0.333	0.333
	SVMRAD	0.455	0.405	0.389	0.372	0.289	0.328	0.333	0.333	0.333	0.333
Distancia 2	SVMTAN	0.407	0.396	0.388	0.380	0.347	0.348	0.333	0.333	0.333	0.333
	RL	0.311	0.306	0.291	0.245	0.142	0.275	0.282	0.250	0.220	0.117
	SVMLIN	0.323	0.309	0.289	0.246	0.143	0.276	0.271	0.250	0.217	0.117
	SVMPOLY	0.469	0.472	0.355	0.447	0.423	0.312	0.325	0.304	0.303	0.282
	SVMRAD	0.316	0.305	0.294	0.252	0.171	0.284	0.276	0.264	0.226	0.122
Distancia 3	SVMTAN	0.316	0.305	0.290	0.253	0.188	0.288	0.276	0.299	0.287	0.181
	RL	0.157	0.151	0.136	0.080	0.015	0.146	0.136	0.114	0.077	0.011
	SVMLIN	0.158	0.148	0.127	0.081	0.013	0.154	0.136	0.114	0.076	0.010
	SVMPOLY	0.421	0.443	0.334	0.176	0.391	0.333	0.226	0.333	0.201	0.301
	SVMRAD	0.163	0.149	0.127	0.108	0.015	0.147	0.138	0.116	0.076	0.013
Distancia 4	SVMTAN	0.165	0.152	0.125	0.089	0.014	0.148	0.156	0.123	0.080	0.012
	RL	0.107	0.097	0.076	0.042	0.010	0.097	0.093	0.069	0.040	0.003
	SVMLIN	0.105	0.099	0.075	0.041	0.004	0.097	0.101	0.071	0.039	0.002
	SVMPOLY	0.360	0.358	0.246	0.240	0.189	0.290	0.333	0.333	0.166	0.282
	SVMRAD	0.108	0.101	0.079	0.040	0.006	0.097	0.122	0.074	0.040	0.004
SVMTAN	0.115	0.096	0.089	0.041	0.006	0.096	0.179	0.080	0.040	0.004	
Distancia 1	$n_1 = n_2 = 100$						$n_1 = 20, n_2 = 100$				
	RL	0.401	0.397	0.398	0.364	0.284	0.167	0.167	0.168	0.171	0.154
	SVMLIN	0.401	0.397	0.403	0.367	0.287	0.167	0.167	0.167	0.167	0.156
	SVMPOLY	0.458	0.412	0.450	0.462	0.491	0.167	0.167	0.167	0.167	0.166
	SVMRAD	0.404	0.400	0.396	0.368	0.288	0.167	0.167	0.167	0.167	0.165
Distancia 2	SVMTAN	0.401	0.437	0.399	0.373	0.286	0.167	0.167	0.167	0.167	0.175
	RL	0.309	0.304	0.284	0.243	0.127	0.161	0.158	0.153	0.142	0.082
	SVMLIN	0.308	0.307	0.284	0.242	0.127	0.167	0.167	0.167	0.167	0.082
	SVMPOLY	0.388	0.444	0.493	0.433	0.475	0.167	0.167	0.163	0.161	0.147
	SVMRAD	0.314	0.300	0.283	0.244	0.126	0.167	0.167	0.161	0.156	0.089
Distancia 3	SVMTAN	0.311	0.302	0.283	0.243	0.125	0.167	0.167	0.169	0.162	0.118
	RL	0.165	0.154	0.125	0.083	0.015	0.104	0.096	0.085	0.058	0.010
	SVMLIN	0.165	0.150	0.124	0.086	0.011	0.101	0.095	0.083	0.054	0.010
	SVMPOLY	0.420	0.476	0.466	0.453	0.058	0.167	0.167	0.123	0.155	0.146
	SVMRAD	0.159	0.150	0.125	0.084	0.012	0.100	0.097	0.088	0.069	0.014
Distancia 4	SVMTAN	0.158	0.149	0.127	0.084	0.012	0.102	0.117	0.085	0.065	0.014
	RL	0.105	0.103	0.075	0.041	0.003	0.072	0.066	0.050	0.030	0.003
	SVMLIN	0.108	0.099	0.075	0.042	0.002	0.084	0.066	0.051	0.028	0.002
	SVMPOLY	0.246	0.432	0.475	0.092	0.400	0.150	0.080	0.152	0.099	0.167
	SVMRAD	0.105	0.102	0.076	0.044	0.004	0.086	0.065	0.057	0.031	0.004
SVMTAN	0.105	0.104	0.076	0.042	0.009	0.096	0.069	0.058	0.035	0.005	

Tabla 4.10.: Resultados distribución normal bivariada $\Sigma_1 = 2\Sigma_2$.

MÉTODO/ ρ		0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
Distancia 1	$n_1 = n_2 = 20$						$n_1 = 20, n_2 = 50$				
	RL	0.458	0.441	0.423	0.390	0.327	0.300	0.286	0.312	0.323	0.288
	SVMLIN	0.440	0.423	0.427	0.385	0.334	0.286	0.286	0.286	0.286	0.286
	SVMPOLY	0.463	0.481	0.406	0.467	0.560	0.286	0.286	0.286	0.286	0.286
	SVMRAD	0.446	0.420	0.375	0.373	0.388	0.286	0.286	0.286	0.286	0.286
Distancia 2	SVMTAN	0.434	0.429	0.454	0.393	0.401	0.286	0.286	0.286	0.286	0.286
	RL	0.332	0.337	0.312	0.281	0.202	0.299	0.296	0.287	0.265	0.160
	SVMLIN	0.328	0.337	0.316	0.279	0.205	0.286	0.286	0.286	0.286	0.168
	SVMPOLY	0.428	0.449	0.408	0.494	0.420	0.286	0.286	0.286	0.286	0.286
	SVMRAD	0.324	0.324	0.349	0.303	0.232	0.286	0.286	0.286	0.263	0.182
Distancia 3	SVMTAN	0.327	0.330	0.342	0.272	0.257	0.286	0.286	0.286	0.300	0.192
	RL	0.206	0.204	0.173	0.128	0.032	0.200	0.195	0.162	0.116	0.028
	SVMLIN	0.213	0.197	0.188	0.125	0.034	0.196	0.195	0.163	0.120	0.031
	SVMPOLY	0.394	0.457	0.336	0.366	0.396	0.286	0.286	0.251	0.285	0.286
	SVMRAD	0.219	0.195	0.163	0.143	0.038	0.224	0.188	0.156	0.123	0.040
Distancia 4	SVMTAN	0.218	0.194	0.175	0.146	0.031	0.236	0.201	0.177	0.140	0.034
	RL	0.160	0.136	0.114	0.092	0.017	0.140	0.146	0.111	0.091	0.009
	SVMLIN	0.165	0.138	0.116	0.089	0.012	0.140	0.151	0.111	0.091	0.009
	SVMPOLY	0.264	0.357	0.271	0.278	0.356	0.286	0.282	0.285	0.218	0.245
	SVMRAD	0.145	0.135	0.123	0.112	0.012	0.144	0.152	0.113	0.091	0.012
SVMTAN	0.158	0.142	0.125	0.123	0.009	0.178	0.149	0.115	0.094	0.038	
Distancia 1	$n_1 = n_2 = 50$						$n_1 = 50, n_2 = 100$				
	RL	0.440	0.408	0.410	0.391	0.320	0.348	0.347	0.335	0.334	0.321
	SVMLIN	0.426	0.411	0.395	0.378	0.310	0.333	0.333	0.333	0.333	0.333
	SVMPOLY	0.471	0.457	0.434	0.494	0.439	0.333	0.333	0.333	0.333	0.333
	SVMRAD	0.385	0.367	0.359	0.371	0.293	0.333	0.333	0.333	0.333	0.333
Distancia 2	SVMTAN	0.431	0.397	0.396	0.401	0.312	0.333	0.333	0.333	0.333	0.333
	RL	0.341	0.340	0.317	0.279	0.180	0.334	0.334	0.308	0.286	0.176
	SVMLIN	0.343	0.329	0.316	0.276	0.185	0.333	0.333	0.333	0.333	0.174
	SVMPOLY	0.361	0.490	0.422	0.483	0.447	0.337	0.333	0.333	0.333	0.326
	SVMRAD	0.323	0.324	0.297	0.280	0.226	0.312	0.333	0.333	0.333	0.170
Distancia 3	SVMTAN	0.343	0.328	0.314	0.280	0.238	0.338	0.333	0.333	0.333	0.226
	RL	0.200	0.192	0.182	0.120	0.032	0.203	0.195	0.167	0.121	0.028
	SVMLIN	0.205	0.188	0.171	0.123	0.032	0.204	0.198	0.173	0.121	0.028
	SVMPOLY	0.491	0.476	0.265	0.323	0.447	0.332	0.333	0.332	0.333	0.333
	SVMRAD	0.212	0.197	0.163	0.143	0.041	0.191	0.185	0.165	0.118	0.028
Distancia 4	SVMTAN	0.230	0.204	0.166	0.140	0.048	0.205	0.208	0.172	0.124	0.030
	RL	0.151	0.136	0.117	0.072	0.017	0.146	0.146	0.116	0.073	0.011
	SVMLIN	0.146	0.143	0.114	0.073	0.016	0.147	0.148	0.121	0.073	0.010
	SVMPOLY	0.469	0.276	0.337	0.361	0.146	0.333	0.233	0.331	0.221	0.333
	SVMRAD	0.146	0.138	0.114	0.073	0.016	0.144	0.150	0.123	0.070	0.011
SVMTAN	0.148	0.137	0.122	0.074	0.038	0.147	0.149	0.128	0.073	0.012	
Distancia 1	$n_1 = n_2 = 100$						$n_1 = 20, n_2 = 100$				
	RL	0.410	0.404	0.414	0.382	0.318	0.348	0.347	0.335	0.334	0.321
	SVMLIN	0.410	0.397	0.413	0.372	0.321	0.333	0.333	0.333	0.333	0.333
	SVMPOLY	0.390	0.424	0.468	0.462	0.458	0.333	0.333	0.333	0.333	0.333
	SVMRAD	0.366	0.364	0.362	0.344	0.291	0.333	0.333	0.333	0.333	0.333
Distancia 2	SVMTAN	0.407	0.436	0.408	0.392	0.309	0.333	0.333	0.333	0.333	0.333
	RL	0.337	0.335	0.316	0.282	0.170	0.334	0.334	0.308	0.286	0.176
	SVMLIN	0.330	0.333	0.311	0.274	0.168	0.333	0.333	0.333	0.333	0.174
	SVMPOLY	0.405	0.420	0.336	0.366	0.411	0.337	0.333	0.333	0.333	0.326
	SVMRAD	0.311	0.303	0.290	0.264	0.162	0.312	0.333	0.333	0.333	0.170
Distancia 3	SVMTAN	0.338	0.327	0.376	0.367	0.168	0.338	0.333	0.333	0.333	0.226
	RL	0.200	0.200	0.168	0.123	0.031	0.203	0.195	0.167	0.121	0.028
	SVMLIN	0.204	0.192	0.168	0.130	0.030	0.204	0.198	0.173	0.121	0.028
	SVMPOLY	0.483	0.355	0.248	0.399	0.474	0.332	0.333	0.332	0.333	0.333
	SVMRAD	0.192	0.181	0.161	0.126	0.028	0.191	0.185	0.165	0.118	0.028
Distancia 4	SVMTAN	0.202	0.188	0.168	0.125	0.030	0.205	0.208	0.172	0.124	0.030
	RL	0.148	0.147	0.116	0.074	0.009	0.146	0.146	0.116	0.073	0.011
	SVMLIN	0.149	0.142	0.115	0.076	0.012	0.147	0.148	0.121	0.073	0.010
	SVMPOLY	0.430	0.255	0.339	0.447	0.297	0.333	0.233	0.331	0.221	0.333
	SVMRAD	0.144	0.138	0.113	0.091	0.018	0.144	0.150	0.123	0.070	0.011
SVMTAN	0.147	0.145	0.115	0.108	0.021	0.147	0.149	0.128	0.073	0.012	

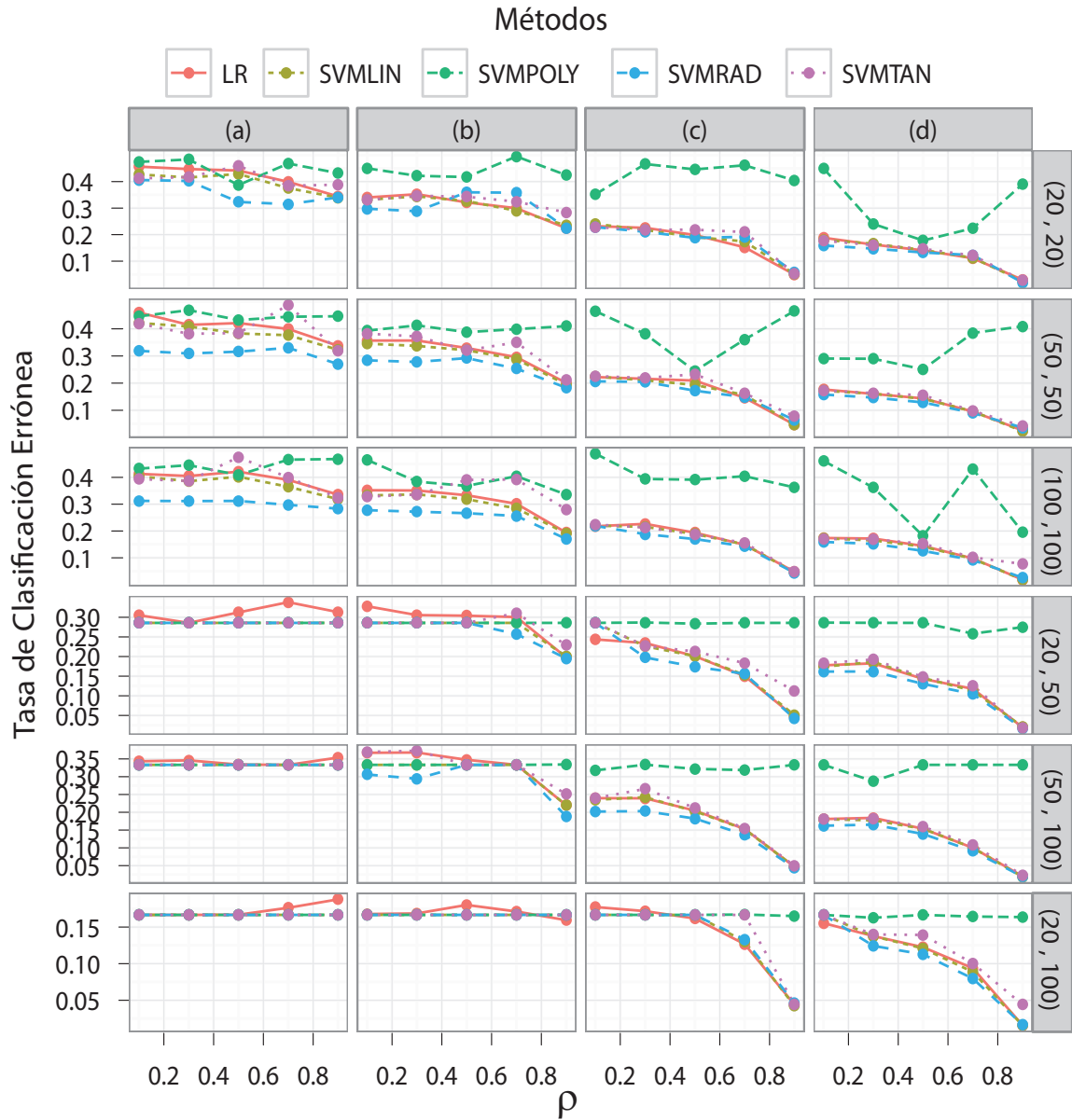


Figura 4.11.: Distribución Normal Bivariada $\Sigma_1 = 3\Sigma_2$

Tabla 4.11.: Resultados distribución Normal Bivariada $\Sigma_1 = 3\Sigma_2$.

MÉTODO/ ρ		0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
Distancia 1	$n_1 = n_2 = 20$						$n_1 = 20, n_2 = 50$				
	RL	0.457	0.447	0.442	0.399	0.344	0.305	0.286	0.312	0.338	0.313
	SVMLIN	0.427	0.416	0.429	0.376	0.340	0.286	0.286	0.286	0.286	0.286
	SVMPOLY	0.475	0.484	0.387	0.468	0.432	0.286	0.286	0.286	0.286	0.286
	SVMRAD	0.406	0.403	0.323	0.315	0.340	0.286	0.286	0.286	0.286	0.286
Distancia 2	SVMTAN	0.412	0.417	0.460	0.385	0.388	0.286	0.286	0.286	0.286	0.286
	RL	0.340	0.352	0.322	0.300	0.224	0.328	0.306	0.304	0.301	0.198
	SVMLIN	0.332	0.344	0.327	0.290	0.235	0.286	0.286	0.286	0.286	0.200
	SVMPOLY	0.450	0.422	0.417	0.495	0.425	0.286	0.286	0.286	0.286	0.286
	SVMRAD	0.298	0.288	0.360	0.358	0.224	0.286	0.286	0.286	0.257	0.195
Distancia 3	SVMTAN	0.331	0.347	0.343	0.324	0.284	0.286	0.286	0.286	0.311	0.229
	RL	0.232	0.226	0.199	0.152	0.049	0.244	0.234	0.201	0.150	0.047
	SVMLIN	0.241	0.218	0.191	0.174	0.051	0.286	0.226	0.201	0.153	0.051
	SVMPOLY	0.352	0.466	0.446	0.462	0.404	0.286	0.287	0.284	0.286	0.286
	SVMRAD	0.228	0.211	0.188	0.192	0.057	0.286	0.198	0.174	0.156	0.042
Distancia 4	SVMTAN	0.228	0.219	0.218	0.210	0.052	0.286	0.227	0.213	0.183	0.112
	RL	0.188	0.163	0.141	0.111	0.029	0.178	0.183	0.143	0.118	0.019
	SVMLIN	0.180	0.166	0.142	0.113	0.025	0.174	0.186	0.145	0.113	0.021
	SVMPOLY	0.450	0.240	0.178	0.224	0.390	0.286	0.286	0.286	0.258	0.275
	SVMRAD	0.159	0.147	0.133	0.123	0.019	0.162	0.162	0.130	0.105	0.017
Distancia 1	SVMTAN	0.177	0.160	0.147	0.121	0.028	0.183	0.192	0.148	0.126	0.018
	$n_1 = n_2 = 50$						$n_1 = 50, n_2 = 100$				
	RL	0.460	0.415	0.421	0.400	0.337	0.343	0.346	0.334	0.333	0.354
	SVMLIN	0.421	0.409	0.383	0.377	0.323	0.333	0.333	0.333	0.333	0.333
	SVMPOLY	0.446	0.469	0.433	0.445	0.446	0.333	0.333	0.333	0.333	0.333
Distancia 2	SVMRAD	0.319	0.309	0.316	0.330	0.270	0.333	0.333	0.333	0.333	0.333
	SVMTAN	0.419	0.382	0.384	0.488	0.319	0.333	0.333	0.333	0.333	0.333
	RL	0.357	0.357	0.329	0.296	0.201	0.367	0.368	0.347	0.333	0.220
	SVMLIN	0.345	0.337	0.321	0.288	0.195	0.333	0.333	0.333	0.333	0.220
	SVMPOLY	0.393	0.413	0.388	0.399	0.410	0.333	0.333	0.333	0.333	0.334
Distancia 3	SVMRAD	0.284	0.278	0.292	0.254	0.183	0.306	0.294	0.333	0.333	0.188
	SVMTAN	0.382	0.372	0.321	0.350	0.212	0.370	0.373	0.333	0.333	0.251
	RL	0.223	0.216	0.209	0.146	0.047	0.240	0.239	0.205	0.153	0.047
	SVMLIN	0.224	0.211	0.193	0.157	0.046	0.235	0.242	0.203	0.152	0.047
	SVMPOLY	0.465	0.381	0.245	0.361	0.466	0.318	0.334	0.322	0.319	0.333
Distancia 4	SVMRAD	0.207	0.205	0.172	0.148	0.064	0.202	0.203	0.182	0.137	0.044
	SVMTAN	0.226	0.219	0.232	0.162	0.078	0.240	0.266	0.212	0.155	0.049
	RL	0.177	0.160	0.144	0.095	0.026	0.181	0.184	0.153	0.100	0.020
	SVMLIN	0.172	0.161	0.143	0.096	0.023	0.180	0.177	0.154	0.100	0.018
	SVMPOLY	0.291	0.290	0.251	0.385	0.409	0.333	0.287	0.333	0.333	0.333
Distancia 1	SVMRAD	0.158	0.147	0.129	0.090	0.036	0.162	0.165	0.139	0.091	0.019
	SVMTAN	0.173	0.163	0.155	0.098	0.042	0.181	0.182	0.160	0.108	0.023
	$n_1 = n_2 = 100$						$n_1 = 20, n_2 = 100$				
	RL	0.413	0.405	0.421	0.391	0.335	0.167	0.167	0.167	0.176	0.188
	SVMLIN	0.404	0.386	0.402	0.365	0.319	0.167	0.167	0.167	0.167	0.167
Distancia 2	SVMPOLY	0.432	0.445	0.409	0.466	0.468	0.167	0.167	0.167	0.167	0.167
	SVMRAD	0.312	0.312	0.312	0.298	0.283	0.167	0.167	0.167	0.167	0.167
	SVMTAN	0.394	0.387	0.475	0.399	0.324	0.167	0.167	0.167	0.167	0.167
	RL	0.352	0.351	0.333	0.302	0.194	0.168	0.169	0.180	0.171	0.159
	SVMLIN	0.333	0.336	0.319	0.284	0.191	0.167	0.167	0.167	0.167	0.167
Distancia 3	SVMPOLY	0.465	0.384	0.368	0.404	0.336	0.167	0.167	0.167	0.167	0.167
	SVMRAD	0.278	0.272	0.266	0.256	0.170	0.167	0.167	0.167	0.167	0.167
	SVMTAN	0.329	0.336	0.390	0.392	0.281	0.167	0.167	0.167	0.167	0.167
	RL	0.218	0.226	0.194	0.149	0.048	0.177	0.172	0.162	0.127	0.044
	SVMLIN	0.220	0.215	0.189	0.150	0.045	0.167	0.167	0.167	0.130	0.042
Distancia 4	SVMPOLY	0.488	0.394	0.391	0.404	0.363	0.167	0.167	0.167	0.167	0.165
	SVMRAD	0.219	0.188	0.170	0.143	0.044	0.167	0.167	0.167	0.133	0.046
	SVMTAN	0.224	0.213	0.187	0.156	0.049	0.167	0.167	0.167	0.167	0.044
	RL	0.174	0.172	0.144	0.098	0.018	0.155	0.138	0.122	0.094	0.016
	SVMLIN	0.173	0.165	0.142	0.096	0.019	0.167	0.137	0.121	0.089	0.015
Distancia 1	SVMPOLY	0.461	0.362	0.182	0.431	0.196	0.167	0.163	0.167	0.164	0.164
	SVMRAD	0.159	0.152	0.126	0.092	0.026	0.167	0.124	0.113	0.080	0.016
	SVMTAN	0.171	0.169	0.152	0.102	0.077	0.167	0.140	0.139	0.100	0.044

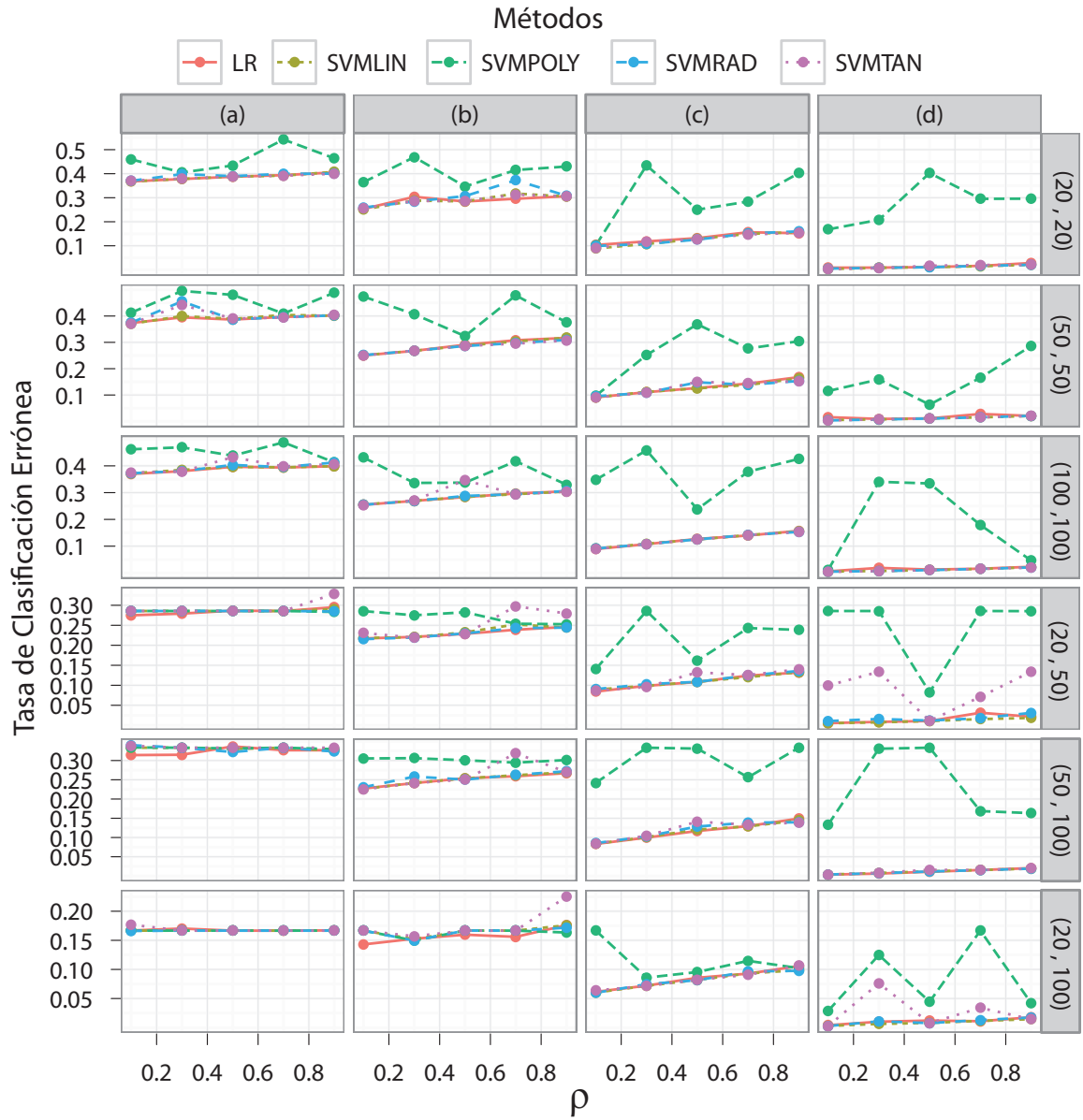


Figura 4.12.: Distribución Normal Multivariada ($p = 200$) $\Sigma_1 = \Sigma_2$

Tabla 4.12.: Resultados distribución Normal Multivariada ($p = 200$) $\Sigma_1 = \Sigma_2$.

		MÉTODO/ ρ	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9	
Distancia 1	Distancia 1	$n_1 = n_2 = 20$						$n_1 = 20, n_2 = 50$					
		RL	0.368	0.378	0.387	0.394	0.405	0.275	0.279	0.287	0.286	0.295	
		SVMLIN	0.368	0.378	0.388	0.392	0.407	0.286	0.286	0.286	0.286	0.290	
		SVMPOLY	0.460	0.405	0.434	0.542	0.464	0.286	0.286	0.286	0.286	0.283	
		SVMRAD	0.370	0.398	0.390	0.400	0.400	0.286	0.286	0.286	0.286	0.285	
Distancia 2	Distancia 2	SVMTAN	0.371	0.380	0.388	0.392	0.400	0.286	0.286	0.286	0.286	0.329	
		RL	0.217	0.221	0.229	0.239	0.246	0.217	0.221	0.229	0.239	0.246	
		SVMLIN	0.218	0.220	0.232	0.252	0.244	0.218	0.220	0.232	0.252	0.244	
		SVMPOLY	0.285	0.275	0.283	0.254	0.253	0.285	0.275	0.283	0.254	0.253	
		SVMRAD	0.215	0.220	0.230	0.243	0.244	0.215	0.220	0.230	0.243	0.244	
Distancia 3	Distancia 3	SVMTAN	0.231	0.219	0.228	0.297	0.280	0.231	0.219	0.228	0.297	0.280	
		RL	0.084	0.099	0.108	0.124	0.132	0.005	0.009	0.011	0.031	0.021	
		SVMLIN	0.089	0.099	0.109	0.120	0.133	0.005	0.007	0.011	0.015	0.018	
		SVMPOLY	0.141	0.286	0.162	0.243	0.238	0.286	0.286	0.082	0.286	0.285	
		SVMRAD	0.090	0.103	0.108	0.125	0.136	0.010	0.016	0.012	0.018	0.031	
Distancia 4	Distancia 4	SVMTAN	0.087	0.096	0.132	0.126	0.140	0.099	0.134	0.011	0.071	0.134	
		RL	0.005	0.009	0.011	0.031	0.021	0.005	0.009	0.011	0.031	0.021	
		SVMLIN	0.005	0.007	0.011	0.015	0.018	0.005	0.007	0.011	0.015	0.018	
		SVMPOLY	0.286	0.286	0.082	0.286	0.285	0.286	0.286	0.082	0.286	0.285	
		SVMRAD	0.010	0.016	0.012	0.018	0.031	0.010	0.016	0.012	0.018	0.031	
Distancia 1	Distancia 1	SVMTAN	0.099	0.134	0.011	0.071	0.134	0.099	0.134	0.011	0.071	0.134	
		$n_1 = n_2 = 50$						$n_1 = 50, n_2 = 100$					
		RL	0.374	0.395	0.387	0.397	0.402	0.314	0.315	0.337	0.327	0.327	
		SVMLIN	0.371	0.400	0.389	0.404	0.401	0.333	0.333	0.330	0.333	0.325	
		SVMPOLY	0.413	0.495	0.480	0.409	0.488	0.334	0.333	0.331	0.333	0.330	
Distancia 2	Distancia 2	SVMRAD	0.375	0.456	0.387	0.396	0.402	0.340	0.333	0.322	0.333	0.323	
		SVMTAN	0.371	0.442	0.390	0.394	0.404	0.338	0.333	0.333	0.333	0.333	
		RL	0.250	0.268	0.291	0.308	0.317	0.227	0.241	0.254	0.259	0.267	
		SVMLIN	0.251	0.268	0.287	0.303	0.317	0.226	0.242	0.254	0.262	0.269	
		SVMPOLY	0.474	0.407	0.324	0.478	0.376	0.305	0.306	0.301	0.295	0.301	
Distancia 3	Distancia 3	SVMRAD	0.251	0.268	0.286	0.297	0.309	0.231	0.258	0.250	0.263	0.272	
		SVMTAN	0.250	0.268	0.286	0.295	0.307	0.225	0.241	0.250	0.319	0.270	
		RL	0.091	0.111	0.127	0.143	0.168	0.004	0.006	0.011	0.016	0.020	
		SVMLIN	0.091	0.112	0.125	0.138	0.163	0.004	0.008	0.012	0.016	0.019	
		SVMPOLY	0.097	0.252	0.369	0.277	0.304	0.133	0.331	0.333	0.169	0.163	
Distancia 4	Distancia 4	SVMRAD	0.095	0.108	0.148	0.139	0.153	0.004	0.007	0.012	0.015	0.019	
		SVMTAN	0.090	0.110	0.150	0.145	0.153	0.003	0.008	0.016	0.015	0.020	
		RL	0.016	0.009	0.011	0.028	0.021	0.004	0.006	0.011	0.016	0.020	
		SVMLIN	0.005	0.007	0.011	0.015	0.020	0.004	0.008	0.012	0.016	0.019	
		SVMPOLY	0.115	0.159	0.064	0.166	0.286	0.133	0.331	0.333	0.169	0.163	
Distancia 1	Distancia 1	SVMRAD	0.004	0.007	0.011	0.015	0.021	0.004	0.007	0.012	0.015	0.019	
		SVMTAN	0.003	0.008	0.012	0.015	0.021	0.003	0.008	0.016	0.015	0.020	
		$n_1 = n_2 = 100$						$n_1 = 20, n_2 = 100$					
		RL	0.370	0.380	0.397	0.394	0.399	0.167	0.170	0.167	0.167	0.167	
		SVMLIN	0.372	0.385	0.396	0.395	0.399	0.167	0.167	0.167	0.167	0.167	
Distancia 2	Distancia 2	SVMPOLY	0.462	0.469	0.438	0.488	0.413	0.167	0.167	0.167	0.167	0.167	
		SVMRAD	0.372	0.380	0.403	0.395	0.413	0.166	0.167	0.167	0.167	0.167	
		SVMTAN	0.375	0.379	0.432	0.398	0.406	0.177	0.167	0.167	0.167	0.167	
		RL	0.255	0.269	0.283	0.297	0.305	0.143	0.153	0.160	0.156	0.176	
		SVMLIN	0.256	0.269	0.284	0.294	0.304	0.167	0.152	0.167	0.167	0.176	
Distancia 3	Distancia 3	SVMPOLY	0.431	0.336	0.337	0.417	0.330	0.167	0.150	0.167	0.167	0.163	
		SVMRAD	0.255	0.269	0.288	0.294	0.306	0.167	0.150	0.167	0.167	0.172	
		SVMTAN	0.254	0.271	0.347	0.294	0.304	0.167	0.157	0.167	0.167	0.225	
		RL	0.089	0.108	0.127	0.140	0.157	0.004	0.010	0.012	0.011	0.018	
		SVMLIN	0.091	0.109	0.125	0.141	0.155	0.003	0.006	0.008	0.012	0.014	
Distancia 4	Distancia 4	SVMPOLY	0.348	0.458	0.238	0.378	0.426	0.029	0.125	0.044	0.167	0.042	
		SVMRAD	0.091	0.108	0.125	0.142	0.153	0.003	0.011	0.008	0.013	0.017	
		SVMTAN	0.089	0.108	0.125	0.142	0.153	0.003	0.076	0.008	0.034	0.014	
		RL	0.006	0.019	0.013	0.016	0.023	0.004	0.010	0.012	0.011	0.018	
		SVMLIN	0.005	0.007	0.011	0.015	0.021	0.003	0.006	0.008	0.012	0.014	
Distancia 1	Distancia 1	SVMPOLY	0.011	0.340	0.334	0.179	0.047	0.029	0.125	0.044	0.167	0.042	
		SVMRAD	0.005	0.007	0.011	0.015	0.021	0.003	0.011	0.008	0.013	0.017	
		SVMTAN	0.004	0.007	0.011	0.015	0.020	0.003	0.076	0.008	0.034	0.014	

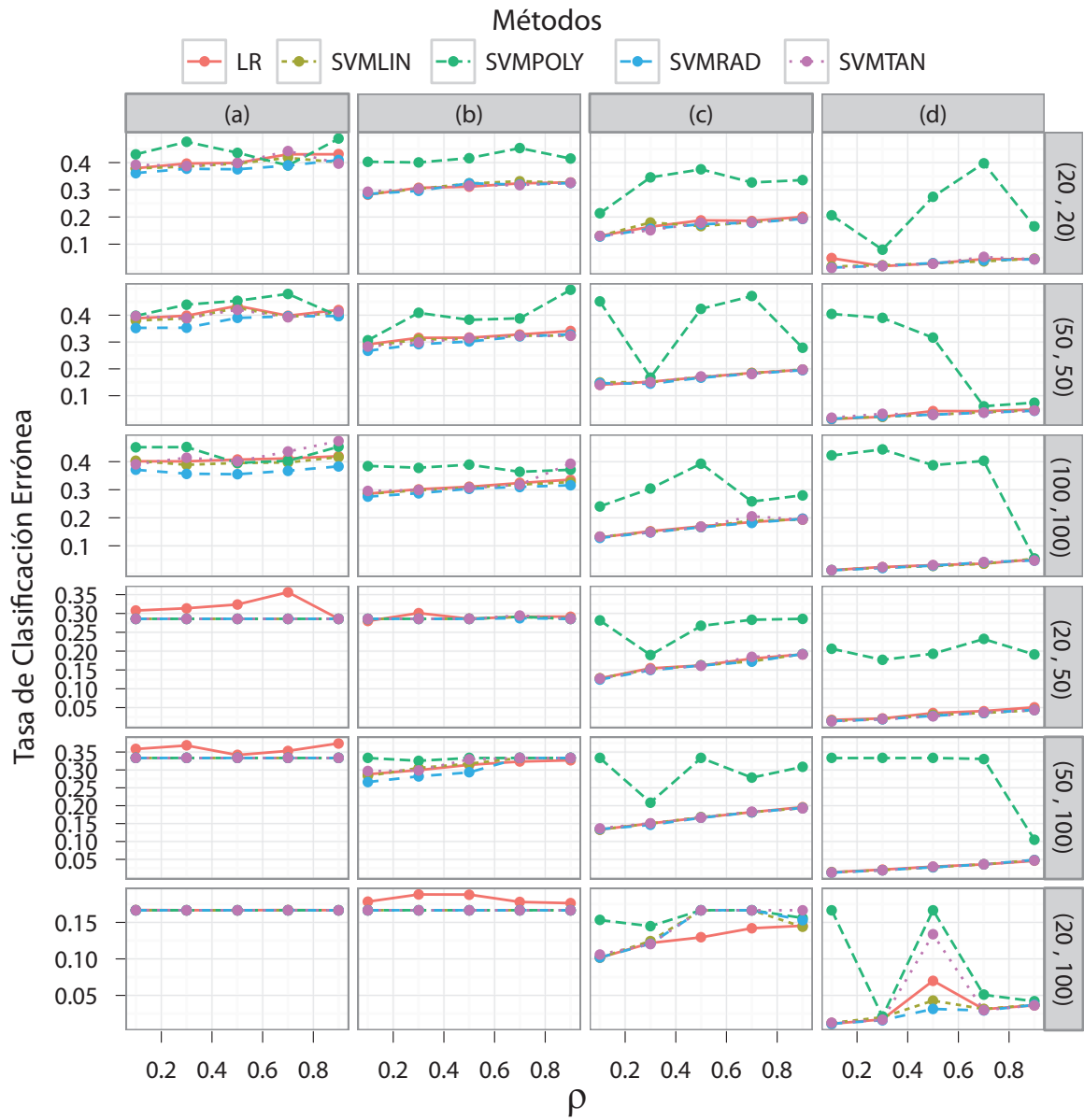


Figura 4.13.: Distribución Normal Multivariada ($p = 200$) $\Sigma_1 = 2\Sigma_2$.

Tabla 4.13.: Resultados distribución Normal Multivariada ($p = 200$) $\Sigma_1 = 2\Sigma_2$.

		MÉTODO/ ρ	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
Distancia 1	$n_1 = n_2 = 20$											
	RL	0.379	0.397	0.399	0.430	0.431	0.308	0.314	0.324	0.357	0.286	0.286
	SVMLIN	0.379	0.386	0.397	0.418	0.401	0.286	0.286	0.286	0.286	0.286	0.286
	SVMPOLY	0.431	0.476	0.436	0.389	0.488	0.286	0.286	0.286	0.286	0.286	0.286
	SVMRAD	0.362	0.377	0.375	0.389	0.409	0.286	0.286	0.286	0.286	0.286	0.286
Distancia 2	$n_1 = 20, n_2 = 50$											
	RL	0.280	0.301	0.286	0.291	0.292	0.280	0.301	0.286	0.291	0.292	0.292
	SVMLIN	0.286	0.286	0.286	0.291	0.286	0.286	0.286	0.286	0.286	0.291	0.286
	SVMPOLY	0.286	0.286	0.286	0.291	0.286	0.286	0.286	0.286	0.286	0.291	0.286
	SVMRAD	0.286	0.286	0.286	0.287	0.286	0.286	0.286	0.286	0.287	0.286	0.286
Distancia 3	$n_1 = n_2 = 50$											
	RL	0.128	0.154	0.162	0.180	0.191	0.017	0.021	0.035	0.041	0.051	0.051
	SVMLIN	0.128	0.152	0.162	0.173	0.192	0.014	0.020	0.030	0.036	0.043	0.043
	SVMPOLY	0.281	0.190	0.267	0.283	0.286	0.206	0.177	0.193	0.232	0.191	0.191
	SVMRAD	0.124	0.149	0.162	0.172	0.192	0.014	0.019	0.028	0.036	0.044	0.044
Distancia 4	$n_1 = 50, n_2 = 100$											
	RL	0.017	0.021	0.035	0.041	0.051	0.017	0.021	0.035	0.041	0.051	0.051
	SVMLIN	0.014	0.020	0.030	0.036	0.043	0.014	0.020	0.030	0.036	0.043	0.043
	SVMPOLY	0.206	0.177	0.193	0.232	0.191	0.206	0.177	0.193	0.232	0.191	0.191
	SVMRAD	0.014	0.019	0.028	0.036	0.044	0.014	0.019	0.028	0.036	0.044	0.044
Distancia 1	$n_1 = n_2 = 100$											
	RL	0.388	0.398	0.435	0.398	0.420	0.358	0.368	0.342	0.353	0.374	0.374
	SVMLIN	0.380	0.388	0.427	0.393	0.410	0.333	0.333	0.333	0.333	0.333	0.333
	SVMPOLY	0.398	0.440	0.454	0.479	0.396	0.333	0.333	0.333	0.333	0.333	0.333
	SVMRAD	0.353	0.353	0.390	0.396	0.397	0.333	0.333	0.333	0.333	0.333	0.333
Distancia 2	$n_1 = 20, n_2 = 100$											
	RL	0.291	0.316	0.316	0.328	0.341	0.288	0.299	0.314	0.323	0.327	0.327
	SVMLIN	0.284	0.309	0.312	0.323	0.324	0.283	0.306	0.317	0.333	0.333	0.333
	SVMPOLY	0.307	0.409	0.383	0.389	0.494	0.333	0.326	0.333	0.333	0.333	0.333
	SVMRAD	0.267	0.292	0.302	0.321	0.328	0.266	0.282	0.293	0.333	0.333	0.333
Distancia 3	$n_1 = n_2 = 20, n_2 = 100$											
	RL	0.141	0.152	0.169	0.185	0.196	0.014	0.021	0.030	0.036	0.046	0.046
	SVMLIN	0.150	0.150	0.170	0.184	0.198	0.014	0.020	0.028	0.036	0.048	0.048
	SVMPOLY	0.451	0.168	0.424	0.471	0.279	0.333	0.333	0.333	0.331	0.105	0.105
	SVMRAD	0.147	0.144	0.167	0.182	0.195	0.013	0.020	0.028	0.036	0.049	0.049
Distancia 4	$n_1 = 20, n_2 = 100$											
	RL	0.013	0.021	0.043	0.042	0.048	0.014	0.021	0.030	0.036	0.046	0.046
	SVMLIN	0.015	0.022	0.029	0.039	0.046	0.014	0.020	0.028	0.036	0.048	0.048
	SVMPOLY	0.405	0.390	0.316	0.061	0.074	0.333	0.333	0.333	0.331	0.105	0.105
	SVMRAD	0.014	0.024	0.029	0.037	0.044	0.013	0.020	0.028	0.036	0.049	0.049
Distancia 1	$n_1 = n_2 = 100$											
	RL	0.401	0.401	0.407	0.411	0.419	0.167	0.167	0.167	0.167	0.167	0.167
	SVMLIN	0.403	0.389	0.395	0.397	0.416	0.167	0.167	0.167	0.167	0.167	0.167
	SVMPOLY	0.452	0.452	0.396	0.403	0.453	0.167	0.167	0.167	0.167	0.167	0.167
	SVMRAD	0.371	0.357	0.355	0.368	0.383	0.167	0.167	0.167	0.167	0.167	0.167
Distancia 2	$n_1 = n_2 = 100$											
	RL	0.391	0.414	0.401	0.436	0.474	0.167	0.167	0.167	0.167	0.167	0.167
	SVMLIN	0.286	0.302	0.311	0.324	0.336	0.178	0.188	0.188	0.178	0.176	0.176
	SVMPOLY	0.285	0.298	0.307	0.318	0.327	0.167	0.167	0.167	0.167	0.167	0.167
	SVMRAD	0.385	0.378	0.389	0.364	0.371	0.167	0.167	0.167	0.167	0.167	0.167
Distancia 3	$n_1 = n_2 = 100$											
	RL	0.275	0.287	0.304	0.310	0.316	0.167	0.167	0.167	0.167	0.167	0.167
	SVMLIN	0.296	0.299	0.307	0.318	0.393	0.167	0.167	0.167	0.167	0.167	0.167
	SVMPOLY	0.130	0.152	0.169	0.185	0.196	0.011	0.017	0.070	0.031	0.036	0.036
	SVMRAD	0.132	0.149	0.167	0.190	0.193	0.012	0.020	0.043	0.031	0.037	0.037
Distancia 4	$n_1 = n_2 = 100$											
	RL	0.240	0.304	0.393	0.258	0.280	0.167	0.021	0.167	0.051	0.042	0.042
	SVMLIN	0.128	0.148	0.166	0.182	0.196	0.011	0.016	0.031	0.029	0.037	0.037
	SVMPOLY	0.133	0.149	0.168	0.205	0.193	0.012	0.016	0.134	0.030	0.036	0.036
	SVMRAD	0.014	0.025	0.031	0.036	0.052	0.011	0.017	0.070	0.031	0.036	0.036
Distancia 1	$n_1 = n_2 = 100$											
	RL	0.013	0.023	0.028	0.036	0.054	0.012	0.020	0.043	0.031	0.037	0.037
	SVMLIN	0.422	0.444	0.387	0.403	0.055	0.167	0.021	0.167	0.051	0.042	0.042
	SVMRAD	0.013	0.021	0.029	0.041	0.048	0.011	0.016	0.031	0.029	0.037	0.037
	SVMRAD	0.013	0.024	0.031	0.042	0.047	0.012	0.016	0.134	0.030	0.036	0.036

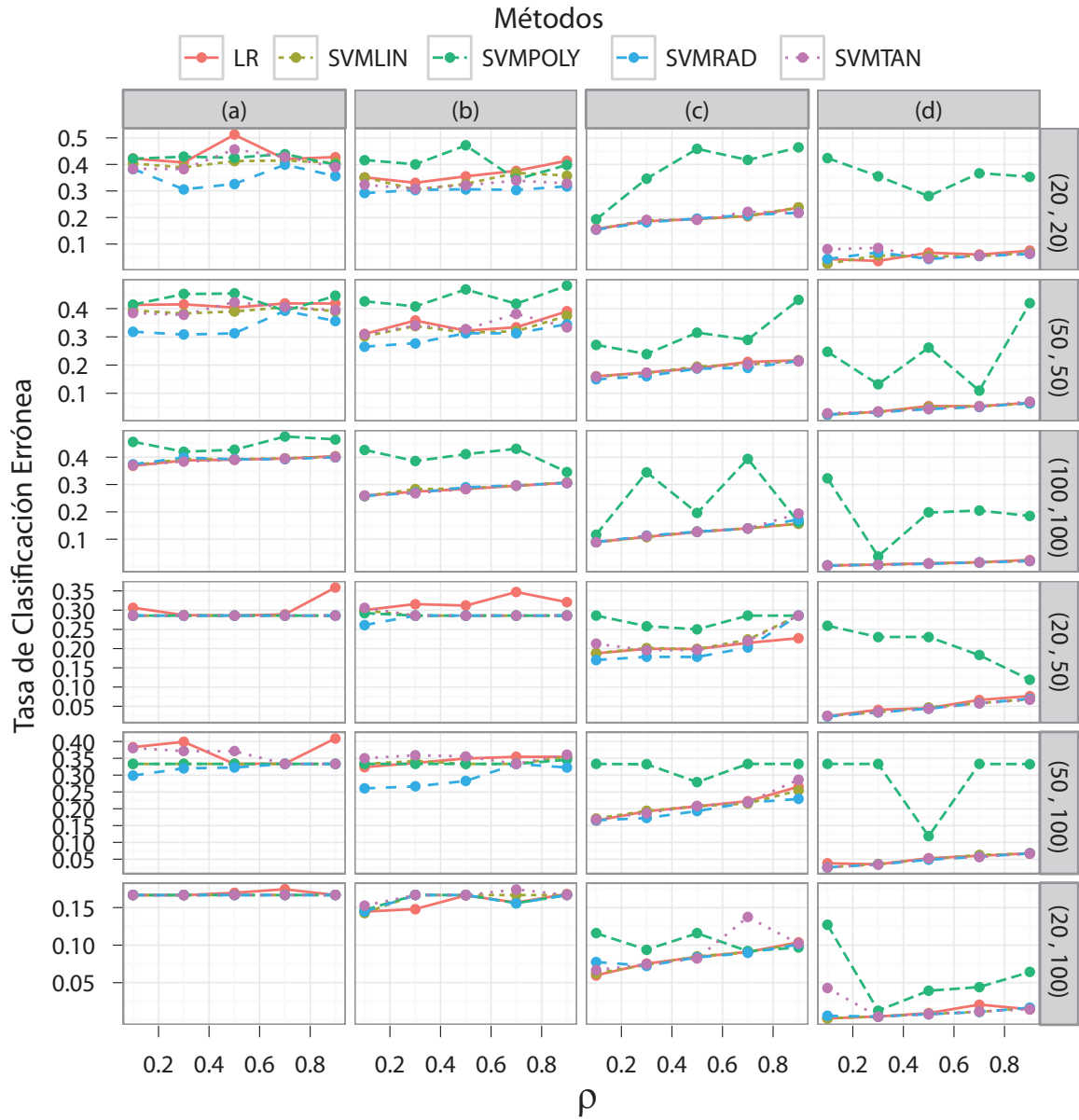


Figura 4.14.: Distribución Normal Multivariada ($p = 200$) $\Sigma_1 = 3\Sigma_2$

Tabla 4.14.: Resultados distribución Normal Multivariada ($p = 200$) $\Sigma_1 = 3\Sigma_2$.

		MÉTODO/ ρ	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
Distancia 1	$n_1 = n_2 = 20$											
	RL	0.422	0.407	0.512	0.420	0.428	0.306	0.287	0.286	0.289	0.359	
	SVMLIN	0.403	0.389	0.412	0.415	0.407	0.286	0.286	0.286	0.286	0.286	
	SVMPOLY	0.422	0.429	0.425	0.439	0.399	0.286	0.286	0.286	0.286	0.286	
	SVMRAD	0.383	0.306	0.327	0.399	0.356	0.286	0.286	0.286	0.286	0.286	
Distancia 2	$n_1 = 20, n_2 = 50$											
	RL	0.300	0.316	0.312	0.347	0.321	0.300	0.316	0.312	0.347	0.321	
	SVMLIN	0.299	0.286	0.286	0.286	0.286	0.299	0.286	0.286	0.286	0.286	
	SVMPOLY	0.292	0.286	0.286	0.286	0.286	0.292	0.286	0.286	0.286	0.286	
	SVMRAD	0.261	0.286	0.286	0.286	0.286	0.261	0.286	0.286	0.286	0.286	
Distancia 3	$n_1 = n_2 = 50$											
	RL	0.188	0.199	0.199	0.215	0.227	0.024	0.041	0.045	0.066	0.077	
	SVMLIN	0.188	0.201	0.199	0.223	0.286	0.025	0.035	0.047	0.058	0.068	
	SVMPOLY	0.286	0.258	0.250	0.286	0.286	0.260	0.230	0.230	0.183	0.119	
	SVMRAD	0.170	0.179	0.178	0.203	0.286	0.023	0.034	0.044	0.058	0.069	
Distancia 4	$n_1 = 50, n_2 = 100$											
	RL	0.213	0.196	0.196	0.220	0.286	0.025	0.035	0.044	0.058	0.067	
	SVMLIN	0.024	0.041	0.045	0.066	0.077	0.024	0.041	0.045	0.066	0.077	
	SVMPOLY	0.025	0.035	0.047	0.058	0.068	0.025	0.035	0.047	0.058	0.068	
	SVMRAD	0.260	0.230	0.230	0.183	0.119	0.260	0.230	0.230	0.183	0.119	
Distancia 1	$n_1 = n_2 = 50$											
	RL	0.023	0.034	0.044	0.058	0.069	0.023	0.034	0.044	0.058	0.069	
	SVMLIN	0.025	0.035	0.044	0.058	0.067	0.025	0.035	0.044	0.058	0.067	
	SVMRAD	0.023	0.034	0.044	0.058	0.069	0.023	0.034	0.044	0.058	0.069	
	SVMRAD	0.023	0.034	0.044	0.058	0.069	0.023	0.034	0.044	0.058	0.069	
Distancia 1	$n_1 = n_2 = 50$											
	RL	0.413	0.415	0.404	0.419	0.418	0.383	0.399	0.333	0.334	0.409	
	SVMLIN	0.394	0.383	0.390	0.405	0.391	0.333	0.333	0.333	0.333	0.333	
	SVMPOLY	0.414	0.452	0.454	0.392	0.445	0.333	0.333	0.333	0.333	0.333	
	SVMRAD	0.318	0.308	0.312	0.393	0.356	0.299	0.320	0.323	0.333	0.333	
Distancia 2	$n_1 = 50, n_2 = 100$											
	RL	0.385	0.378	0.423	0.405	0.394	0.381	0.371	0.371	0.333	0.333	
	SVMLIN	0.310	0.358	0.322	0.334	0.390	0.324	0.336	0.349	0.354	0.354	
	SVMPOLY	0.302	0.338	0.315	0.320	0.375	0.333	0.342	0.333	0.333	0.354	
	SVMRAD	0.426	0.408	0.468	0.417	0.481	0.333	0.333	0.333	0.333	0.346	
Distancia 3	$n_1 = n_2 = 100$											
	RL	0.265	0.277	0.313	0.312	0.345	0.261	0.266	0.283	0.333	0.323	
	SVMLIN	0.308	0.340	0.327	0.381	0.334	0.351	0.359	0.356	0.333	0.361	
	SVMRAD	0.160	0.174	0.189	0.211	0.217	0.038	0.035	0.053	0.060	0.067	
	SVMRAD	0.159	0.172	0.195	0.203	0.215	0.026	0.035	0.050	0.062	0.067	
Distancia 4	$n_1 = 20, n_2 = 100$											
	RL	0.271	0.239	0.315	0.290	0.431	0.333	0.333	0.118	0.333	0.333	
	SVMLIN	0.150	0.161	0.187	0.190	0.214	0.026	0.034	0.048	0.058	0.067	
	SVMRAD	0.158	0.173	0.190	0.202	0.214	0.026	0.035	0.052	0.056	0.067	
	SVMRAD	0.158	0.173	0.190	0.202	0.214	0.026	0.035	0.052	0.056	0.067	
Distancia 1	$n_1 = n_2 = 100$											
	RL	0.025	0.035	0.055	0.054	0.067	0.038	0.035	0.053	0.060	0.067	
	SVMLIN	0.024	0.034	0.051	0.054	0.064	0.026	0.035	0.050	0.062	0.067	
	SVMRAD	0.247	0.132	0.262	0.109	0.419	0.333	0.333	0.118	0.333	0.333	
	SVMRAD	0.025	0.033	0.044	0.052	0.066	0.026	0.034	0.048	0.058	0.067	
Distancia 2	$n_1 = 20, n_2 = 100$											
	RL	0.029	0.035	0.044	0.054	0.071	0.026	0.035	0.052	0.056	0.067	
	SVMLIN	0.369	0.388	0.391	0.395	0.404	0.166	0.166	0.170	0.174	0.167	
	SVMRAD	0.371	0.393	0.393	0.396	0.402	0.167	0.167	0.167	0.167	0.167	
	SVMRAD	0.374	0.399	0.392	0.393	0.400	0.167	0.167	0.167	0.167	0.167	
Distancia 3	$n_1 = 20, n_2 = 100$											
	RL	0.370	0.384	0.390	0.395	0.400	0.167	0.167	0.167	0.167	0.167	
	SVMLIN	0.260	0.274	0.284	0.295	0.307	0.145	0.148	0.166	0.157	0.168	
	SVMRAD	0.258	0.283	0.286	0.297	0.307	0.143	0.167	0.167	0.167	0.167	
	SVMRAD	0.427	0.386	0.411	0.431	0.346	0.147	0.167	0.167	0.156	0.167	
Distancia 4	$n_1 = 20, n_2 = 100$											
	RL	0.258	0.270	0.290	0.297	0.306	0.144	0.167	0.167	0.156	0.167	
	SVMLIN	0.260	0.268	0.283	0.296	0.305	0.153	0.167	0.167	0.174	0.167	
	SVMRAD	0.090	0.109	0.126	0.140	0.156	0.003	0.005	0.009	0.021	0.015	
	SVMRAD	0.089	0.108	0.129	0.140	0.156	0.003	0.005	0.008	0.011	0.016	
Distancia 1	$n_1 = 20, n_2 = 100$											
	RL	0.117	0.345	0.196	0.394	0.162	0.127	0.013	0.039	0.044	0.064	
	SVMLIN	0.089	0.113	0.128	0.139	0.172	0.006	0.005	0.008	0.011	0.017	
	SVMRAD	0.089	0.113	0.128	0.139	0.172	0.006	0.005	0.008	0.011	0.017	
	SVMRAD	0.089	0.110	0.127	0.139	0.193	0.043	0.005	0.008	0.012	0.015	
Distancia 2	$n_1 = 20, n_2 = 100$											
	RL	0.004	0.007	0.012	0.015	0.024	0.003	0.005	0.009	0.021	0.015	
	SVMLIN	0.004	0.007	0.010	0.015	0.021	0.003	0.005	0.008	0.011	0.016	
	SVMRAD	0.322	0.037	0.198	0.205	0.185	0.127	0.013	0.039	0.044	0.064	
	SVMRAD	0.004	0.007	0.010	0.015	0.020	0.006	0.005	0.008	0.011	0.017	
Distancia 3	$n_1 = 20, n_2 = 100$											
	RL	0.003	0.007	0.011	0.015	0.020	0.043	0.005	0.008	0.012	0.015	
	SVMLIN	0.004	0.007	0.010	0.015	0.021	0.003	0.005	0.008	0.011	0.016	
	SVMRAD	0.004	0.007	0.010	0.015	0.020	0.006	0.005	0.008	0.011	0.017	
	SVMRAD	0.003	0.007	0.011	0.015	0.020	0.043	0.005	0.008	0.012	0.015	

5. Aplicaciones Genéticas

En este capítulo se pondrán a prueba el desempeño de los métodos de clasificación considerados en este estudio con datos provenientes de estudios con microarreglos en los que se consideran pacientes con la enfermedad (casos) y personas sanas. Las dos aplicaciones que se presentan corresponden a investigaciones genéticas en pacientes con diabetes tipo 2 y enfermedad de Alzheimer. Se iniciará con definiciones básicas que permitirán entender la naturaleza de los datos y la estrategia empleada para su análisis. Posteriormente se realiza un análisis descriptivo de cada conjunto de datos y finalmente se ajustarán los modelos SVM y de RL y se calculará la MCR.

5.1. Expresión Genética

“Como libros en una biblioteca, el propósito de los genes es almacenar información. Cada gen es un libro que contiene la información requerida para producir una proteína, o en algunos casos un ARN (Ácido Ribonucleico) no codificado. De la misma manera que los libros se puede tomar de un estante y leer, los genes se expresan para producir ARN funcionales y las moléculas de proteínas en la células.”

Twyman (2003)

En los diferentes tipos de células, no todos los genes se expresan de igual manera o al mismo tiempo. Por ello, cuando se realizan estudios tipo caso/control, pueden utilizarse los niveles de expresión genética como variables explicativas para la clasificación de individuos.

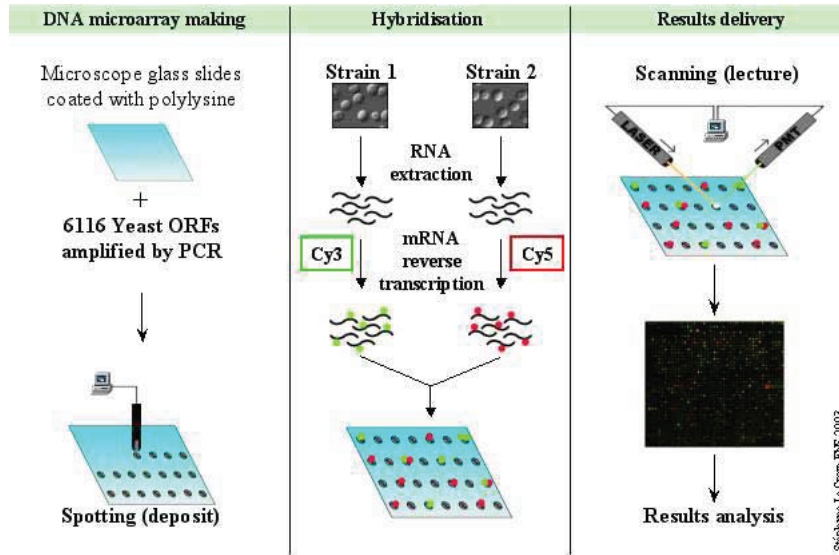


Figura 5.1.: Procedimiento general para obtención de GE.

Específicamente, *Gene Expression* (GE) hace referencia a la medición de los niveles de luminosidad de los colores verde y rojo que se obtienen después someter los genes a complejos procedimientos bioquímicos (Nguyen et al. 2002). En la figura 5.1, tomada de <http://bit.ly/HM2ciN>, se presenta el procedimiento general para de la obtención de los datos.

Como resultado, se obtiene una matriz de $\mathbf{X}_{k \times (n_1 + n_2)}$ donde k es el cantidad de genes, n_1 y n_2 son la cantidad de individuos en los casos y controles, respectivamente, y la ij -ésima entrada representa el nivel de intensidad del gen i en la muestra j ($i = 1, \dots, k; j = 1, \dots, n_1 + n_2$). Para identificar los genes diferencialmente expresados entre casos y controles se usa el estadístico

$$t_i = \frac{\bar{x}_{2i} - \bar{x}_{1i}}{\sqrt{\frac{s_{1i}^2}{n_1} + \frac{s_{2i}^2}{n_2}}} \quad (5.1)$$

donde \bar{x}_{2i} y \bar{x}_{1i} es la intensidad promedio de los casos y controles, respectivamente, mientras que s_{1i}^2 y s_{2i}^2 representan las varianzas muestrales. Valores elevados de t_i indican que el gen i posee niveles diferentes de expresión entre casos y controles. Para llevar a cabo las comparaciones entre los niveles de DE, en las aplicaciones

de las secciones 5.2 y 5.3 se realizarán *hipótesis múltiples* basadas en el estadístico 5.1. En concreto se analizarán simultáneamente miles de pruebas de hipótesis independientes (una para cada gen), como se menciona en Vélez (2008) y Correa (2010) para este tipo de pruebas debe controlarse la proporción de genes incorrectamente clasificados como “expresados”. Sin embargo, por tratarse de simples ilustraciones no se realizará ningún tipo de corrección sobre los valores p en ninguna de las dos aplicaciones, consideradas en este estudio.

5.2. Datos sobre Diabetes

Mootha et al. (2003) presenta una estrategia de análisis para la detección de pequeños cambios en los niveles de expresión de grupos de genes relacionados funcionalmente y lo ilustra con microarreglos de ADN. Los autores miden los niveles de expresión génica en biopsias de 43 hombres de la misma edad, 17 de ellos con tolerancia normal a la glucosa (NGT) , 8 con intolerancia a la glucosa (IGT) y 18 con diabetes tipo 2 (DT2). Como resultado, los autores identificaron un conjunto de genes implicados en la fosforilación oxidativa, un proceso metabólico que utiliza energía liberada por la oxidación de nutrientes para producir adenosín trifosfato.

Para el análisis de los niveles de expresión se procedió de la siguiente forma. Primero, se seleccionaron aleatoriamente un conjunto de 1000 genes de la base de datos original. Segundo, el nivel de expresión en la muestras provenientes de individuos con DT2 (casos, grupo 2) e individuos con NGT (controles, grupo 1) fueron comparados usando el estadístico (5.1), implementadas en la librería `genefilter` de R (Gentleman et al. 2011). Tercero, se utilizaron los 30 genes con niveles más altos de expresión para ajustar los modelos RL y SVM. Este número de genes se seleccionó pues en pruebas piloto se observó que son suficientes para lograr una MCR cercana a cero. Los genes se fueron incluyendo en el modelo uno a uno y en cada paso se calculo MCR por medio de validación cruzada interna (ver Anexo B).

Algunas medidas para los 10 genes con mayor diferencia en expresión se presentan

en la tabla 5.1; los niveles de expresión en pacientes con DT2 son más bajos que en las muestras de NGT en los genes G557, G226 y G137 T2D. La figura 5.2 muestra los diagramas de dispersión para los primeros 5 genes por estado de la enfermedad. Allí se observan algunas estructuras de correlación que podrían constituir un problema potencial para algunos métodos de clasificación.

Tabla 5.1.: Estadísticas de 10 genes diferencialmente expresados. No se aplicó corrección por pruebas múltiples sobre los valores p .

Gen	Estadístico t	$\bar{x}_{\text{NGT}} - \bar{x}_{\text{T2D}}$	Valor- p
G557	3.8788	0.1632	0.0005
G591	-3.6406	-0.1008	0.0009
G226	3.0621	0.1285	0.0044
G718	-3.0566	-0.1093	0.0044
G45	-2.8978	-0.1275	0.0066
G137	2.8432	0.1255	0.0076
G737	-2.6544	-0.1947	0.0121
G587	-2.5774	-0.2654	0.0146
G232	-2.5607	-0.3213	0.0152
G185	-2.5368	-0.2752	0.0161

RL y SVM fueron fijados usando el *status* de la enfermedad como variable dependiente y los niveles de expresión de k genes, como covariables. Los resultados se reportan en la figura 5.3. Para predecir el tipo de enfermedad en este conjunto de datos, (i) SVM requiere menos variables (genes); (ii) todos los métodos se comportaron de forma similar cuando $k < 5$, pero el SVM radial fue más estable en términos de MCR, y (iii) SVM polinomial y tangencial definitivamente no son una buena alternativa, pues como se puede observar en la figura 5.3, estos presentan las MCR más altas.

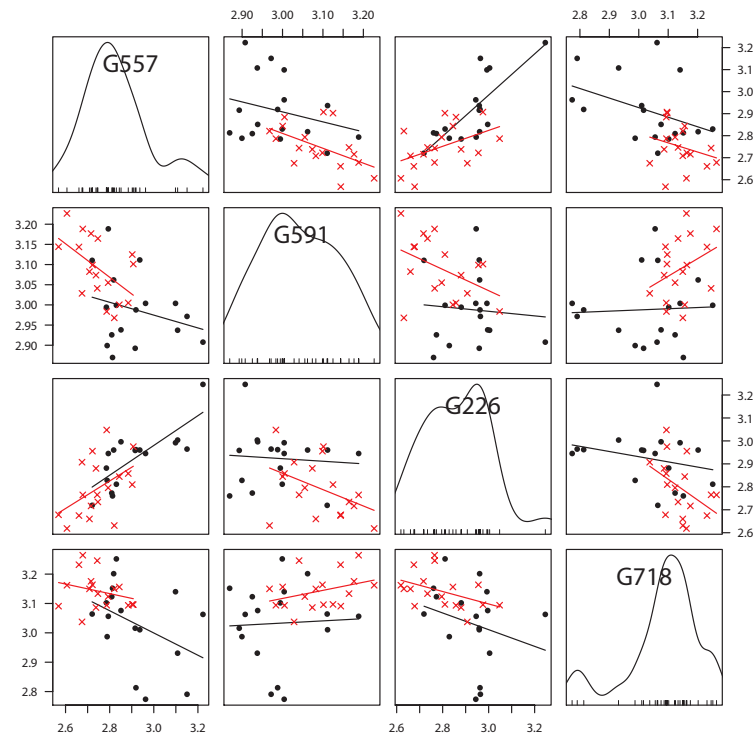


Figura 5.2.: Matriz de diagramas de dispersión para algunos genes presentados en la tabla 5.1. Los puntos representan al grupo NGT (controles); las líneas corresponden a modelos de regresión lineal (controles en negro). En el panel diagonal se muestra el gráfico de densidad.

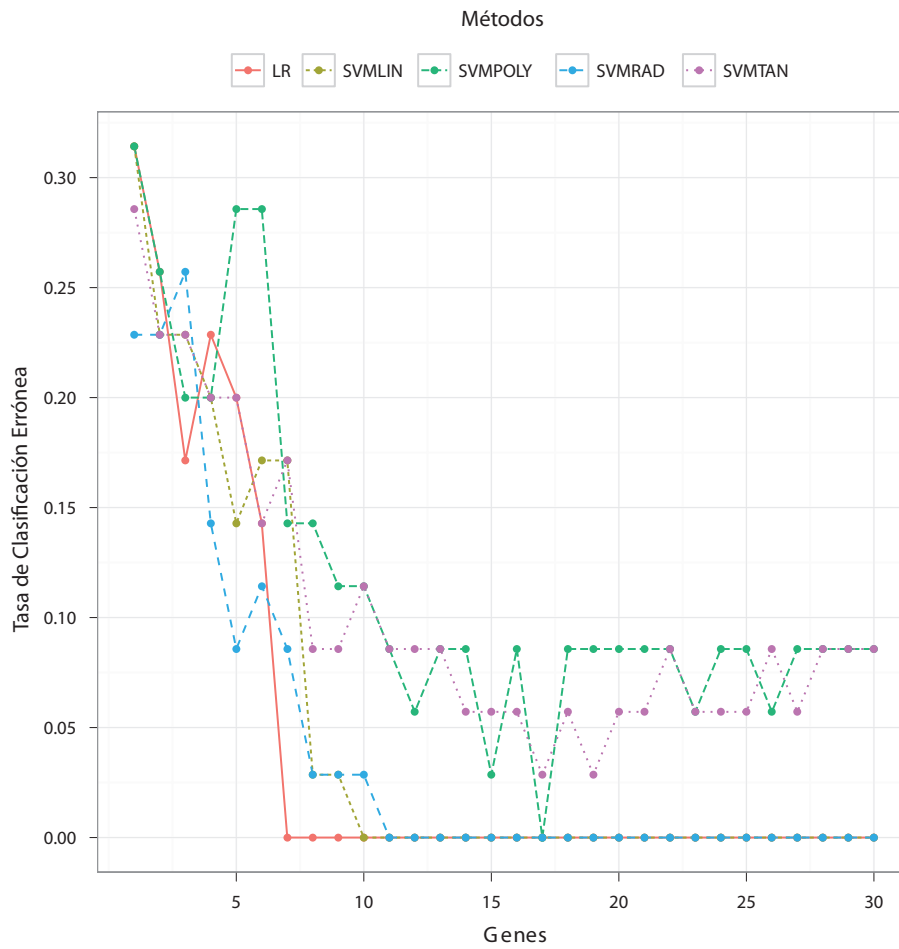


Figura 5.3.: MCR como función del número de genes incluidos en los modelos.

5.3. Datos sobre Alzheimer

La demencia es un término general para describir la pérdida de memoria y de otras habilidades intelectuales. Según la Asociación del Alzheimer de los Estados Unidos, el Alzheimer es una enfermedad neurodegenerativa que representa del 60 % al 80 % de los casos de la demencia (ALZ 2011). Se estima que cerca de 5.4 millones de personas padecen de Alzheimer solo en Estados Unidos y que este número aumente hasta 11-16 millones de personas en el año 2050. Debido a que un 70 % ciento de aquellas personas que padecen de Alzheimer viven en sus hogares, el impacto de esta enfermedad se extiende a millones de familiares y amigos. El centro nacional de información sobre avances biotecnológicos en la ciencia y la salud en Estados Unidos (NCBI 2011), proporciona acceso a información biomédica y genética en diferentes áreas. En GSE26927 (2011) se encuentran los niveles de expresión de 20590 genes en 18 individuos del Reino Unido de los cuales 11 padecen la enfermedad. Se procederá a realizar un análisis similar al de la sección anterior con estos datos para evaluar la capacidad de los métodos para clasificar los enfermos de los sanos.

Tabla 5.2.: Estadísticas de 10 genes diferencialmente expresados. No se aplicó corrección por pruebas múltiples sobre los valores p .

Gen	Estadístico t	$\bar{x}_{\text{Caso}} - \bar{x}_{\text{Control}}$	Valor- p
G856	3.7990	14.7483	0.0016
G927	3.2936	165.5134	0.0046
G64	-3.1401	-156.8164	0.0063
G368	3.0520	16.1379	0.0076
G988	-2.8749	-29.4188	0.0110
G487	2.8181	23.3131	0.0124
G632	-2.7640	-30.0896	0.0138
G591	-2.7113	-51.4538	0.0154
G186	-2.7040	-177.3965	0.0156
G525	-2.6890	-65.2791	0.0161

Luego de seleccionar aleatoriamente 1000 genes de la base original y filtrar los 30

genes predominantes, se obtuvieron los resultados presentados en la tabla 5.2 y las figuras 5.4 y 5.5. RL y SVM fueron ajustados usando el *status* de la enfermedad como variable dependiente y los niveles de expresión de k genes, como covariables.

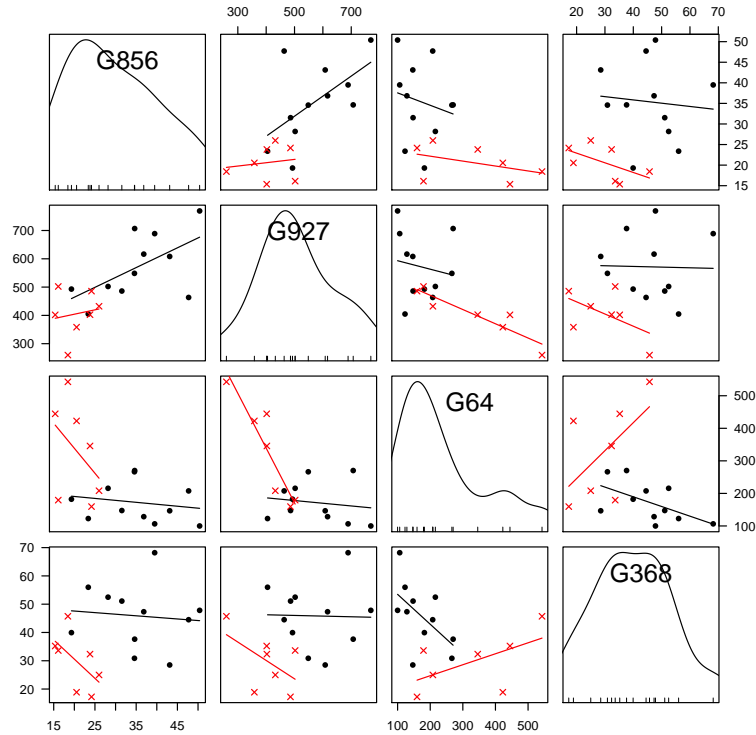


Figura 5.4.: Matriz de diagramas de dispersión para algunos genes presentados en la tabla 5.2. Los puntos representan al grupo NGT (controles); las líneas corresponden a modelos de regresión lineal (controles en negro). En el panel diagonal se muestra el gráfico de densidad.

Al comparar las líneas de tendencia para casos (en negro) y controles (en rojo) de la figura 5.4 con aquellas en la 5.2, nuevamente se evidencian indicios algunas estructuras de correlación entre los niveles de expresión, con menor intensidad para los datos provenientes de la enfermedad del Alzheimer. De la figura 5.5 se puede concluir que para predecir el padecimiento de la enfermedad de Alzheimer en este conjunto de datos (i) SVM requiere menos variables (genes); (ii) los kernels

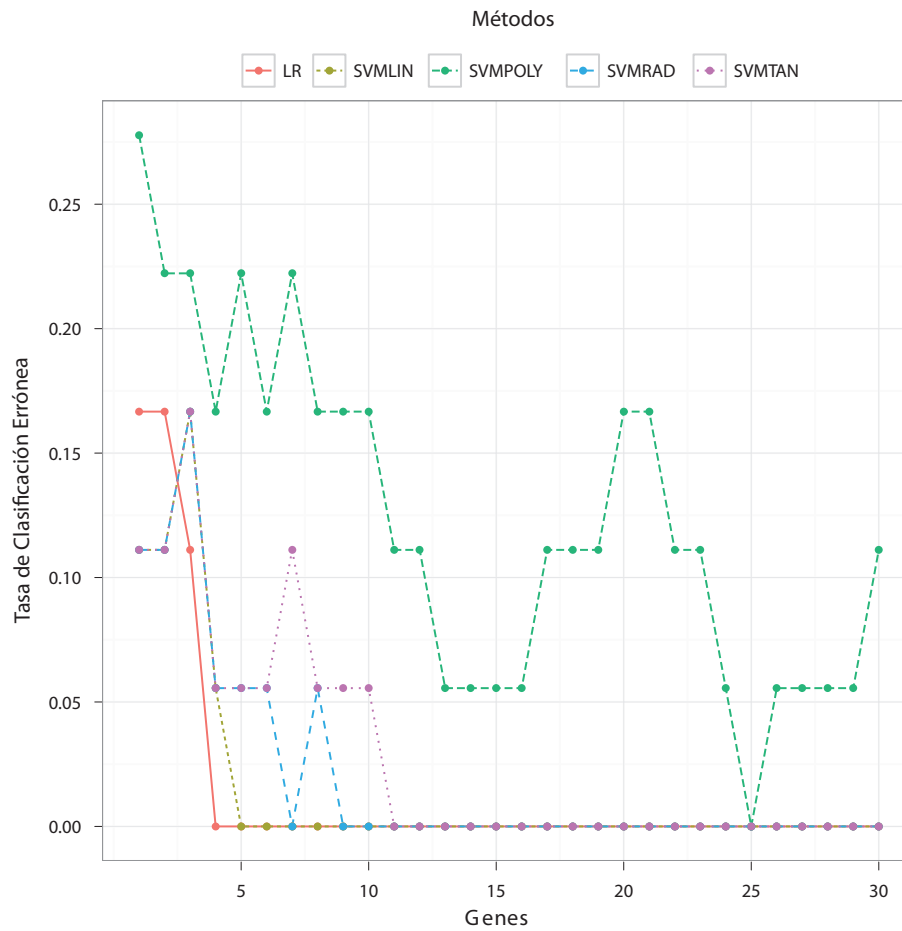


Figura 5.5.: MCR como función del número de genes incluidos en en los modelos.

lineal y radial son una alternativa a considerar, sobre todo cuando $k > 10$ y (iii) SVM polinomial y tangencial definitivamente no es una buena alternativa, ya que no logran estabilizarse en términos de las MCR.

6. Conclusiones Generales

Se logró construir un marco de referencia para la comparación, vía simulación estadística, del desempeño de métodos de clasificación cuando los individuos pertenecen a una de dos categorías mutuamente excluyentes. En particular en este trabajo se comparan las SVM y RL. Con base a los escenarios y simulaciones realizadas se puede concluir que:

1. Cuando el interés es predecir el grupo al que pertenece una *nueva* observación basándose en una sola variable, los modelos SVM son una alternativa viable a RL. Sin embargo, como se muestra en las distribuciones Poisson, Exponencial y Normal, el modelo de SVM polinomial no es recomendable ya que su MCR porcentual es más alta en la mayoría de los escenarios considerados. En el caso de la combinación de distribuciones de probabilidad, a excepción del kernel polinomial, las SVM tienen un mejor desempeño frente a la RL cuando los valores esperados entre los grupos difieren poco.
2. Cuando en el modelo que se quiere utilizar para la clasificación de una *nueva* observación se incluye más de una covariable, se observó una disminución considerable de la MCR de cada método a medida que la correlación entre las covariables aumenta. Respecto al desempeño de cada método, se observó que el kernel polinomial no es una alternativa a considerar. La cantidad de covariables no parece ser un parámetro que afecte considerablemente el desempeño de SVM y RL. El kernel radial y lineal le compiten muy bien a la RL en estas situaciones, pero se mostró que cuando las distancias entre grupos eran pequeñas, y las matrices de variancias y covarianzas eran diferentes, fue más recomendable usar las SVM con un kernel radial.

3. A pesar de que los parámetros de SVM fueron controlados a través de la función `tune.svm()` de R (ver sección 3.3), en algunos escenarios se observaron oscilaciones no esperadas en los valores de la MCR. Como dirección futura se podría indagar más sobre este comportamiento. Se mostró, además, que en la mayoría de los escenarios considerados el kernel polinomial no compite con RL ni con los demás kernels.
4. En cuanto a la aplicación con datos reales, las SVM son una alternativa a considerar. Como se mostró en ambas aplicaciones, las SVM tuvieron un desempeño destacable frente a la RL y requirieron menos covariables para proporcionar una MCR equivalente o mejor a la obtenida con RL. Esto es consistente con lo afirmado en Verplancke et al. (2008). Comparando los resultados de las dos aplicaciones (figuras 5.2 y 5.4) que existe una leve diferencia en la estructura de correlación de los datos. En el gráfico 5.5, se observa una mejoría sustancial del desempeño del kernel tangencial, frente al que se observa 5.3. Con base en estos resultados, se puede concluir que las estructuras de correlación entre las covarianzas afecta el desempeño del kernel tangencial.
5. No solo las SVM tienen un buen desempeño frente a las RL, proporcionando resultados equivalentes, y en algunos casos presentándose como una mejor opción, sino que desde el punto de vista teórico también compiten muy bien. Como se mostró en las primeras secciones, los dos métodos no tienen supuestos teóricos o restricciones fuertes que limiten su implementación. Además, ambos métodos se basan en ideas claras y sencillas que garantizan una solución a los problemas de clasificación.
6. En futuras investigaciones se podría estudiar el comportamiento de los métodos en escenarios complementarios donde se tengan consideraciones adicionales como: grupos con matrices de varianza y covarianza donde una no sea múltiplo de la otra, datos provenientes de distribuciones de probabilidad multivariedades diferentes de la distribución normal, grupos de clasificación que no sean mutuamente excluyentes, es decir, que se consideren grupos espúreos, así como el incluir mas de dos grupos al problema de clasificación.

A. Anexo: Programas en R

A continuación se presenta el código en R utilizado para realizar las simulaciones presentadas en este documento. El algoritmo está compuesto de una función auxiliar y una función principal. En la primera se (`generaD()`) especifica la distribución de probabilidad y se usa para generar los datos de entrenamiento y validación y en la segunda (`ANBEN()`), la cual depende de los parámetros `mu1`, `mu2`, `sigma1`, `sigma2`, `n`, `o`, y `B` los cuales representan en su orden los vectores de medias, matrices de varianzas y covarianzas y cantidad de individuos en cada grupo y por ultimo la cantidad de iteraciones, respectivamente. Como resultado se obtiene la tasa promedio de clasificación errónea para cada método.

```
# ----- #      funcion auxiliar #
----- # funcion para generar D usando una
normal bivariada generaD <- function(mu1, mu2, sigma1, sigma2, n,
o){
  datos1 <- mvrnorm(n, mu1, sigma1)
  datos2 <- mvrnorm(o, mu2, sigma2)
  group <- as.factor(rep(1:0, c(n, o)))
  D1<- c(datos1[, 1], datos2[, 1])
  D2<- c(datos1[, 2], datos2[, 2])
  D <- data.frame(D1, D2, group)
  D
}

# ----- #      funcion principal #
----- ANBEN <- function(mu1, mu2,
sigma1, sigma2, n, o, B = 5000){ # ECR for the SVM models ecrsvm <-
function(modelo, D1val, D2val, data = D){
```

```

    pred <- predict(modelo, data.frame(D1 = D1val, D2 = D2val))
    tab <- table(pred, D[, 3])
    (tab[2,1]+tab[1,2])/sum(tab)
} # MCR para modelo logistico ecrml <- function(modelo, D1val,
D2val, data = D){
    pred <- predict(modelo, data.frame(D1 = D1val, D2 = D2val),
        type = "response")
    out <- data.frame(D[,3], pred, 1-pred)
    out <- cbind(out, predstatus = ifelse(apply(out[,-1], 1,
        which.max) == 1, 1, 0))
    ta <- table(factor(out[,1], levels = 0:1),
        factor(out[,4], levels = 0:1))
    (ta[2,1] + ta[1,2])/sum(ta)
}

# Conjunto de entrenamiento D <- generaD(mu1 = mu1, mu2 = mu2,
sigma1 = sigma1,
    sigma2 = sigma2, n = n, o = o)
status <- D[, 3] # metodos de clasificación tuned <-
tune.svm(group~., data = D, gamma = 10^(-6:-1),
    cost = 10^(-1:1))
cc <- as.numeric(tuned$best.parameters[2]) gg <-
as.numeric(tuned$best.parameters[1]) mylogit <- glm(group ~ ., data
= D, family = binomial) modelolin <- svm(group ~ ., D, type =
"C-classification", cost = cc,
    gamma = gg, kernel = "linear")
modelopoly <- svm(group ~ ., D, type = "C-classification", cost =
cc,
    gamma = gg, kernel = "polynomial")
modelorad <- svm(group ~ ., D, type = "C-classification", cost = cc,
    gamma = gg, kernel = "radial")
modelotan <- svm(group ~ ., D, type = "C-classification", cost = cc,
    gamma = gg, kernel = "sigmoid")

results <- function(mu1, mu2, sigma1, sigma2, n, o){

```

```
# generar nuevos datos
datos1val <- mvrnorm(n, mu1, sigma1)
datos2val <- mvrnorm(o, mu2, sigma2)
D1val <- c(datos1val[,1], datos2val[,1])
D2val <- c(datos1val[,2], datos2val[,2])
# ECR
out <- c(RL = ecrrl(mylogit, D1val, D2val),
        SVMLIN = ecrsvm(modelolin, D1val, D2val),
        SVMPOLY = ecrsvm(modelopoly, D1val, D2val),
        SVMRAD = ecrsvm(modelorad, D1val, D2val),
        SVMTAN = ecrsvm(modelotan, D1val, D2val))
out
} # repitiendo B veces replicate(B, results(mu1, mu2, sigma1,sigma2,
n, o)) }
```

B. Anexo: Algoritmo Aplicaciones

```
# librerias require(bootstrap) require(genefilter) require(e1071)
require(car) require(lattice) require(xtable)

# leyendo los datos d <- as.matrix(read.csv('naturepaper.csv',
header = TRUE)) type <- rep(c('NGT', 'T2D'), c(17,18)) # tipo
biopsias colnames(d) <- paste('s', as.numeric(as.factor(type)),
sep="") tvalues <- abs(rowttests(d, factor(type), tstatOnly =
FALSE)[,1]) o <- order(tvalues, decreasing = TRUE) # order

# Tabla para LaTeX xxx <- rowttests(d, factor(type), tstatOnly =
FALSE) xtable(yyy[order(abs(yyy[, 'statistic']),
decreasing = TRUE), ][1:10,], digits = 4)

genes <- t(d[o[1:5],]) genes2 <- data.frame(type = factor(type),
genes) colnames(genes2)[2:6] <- paste('G', o[1:5], sep = "")

# scatterplotMatrix par(mfrow = c(1,1), mar = c(5, 4, 3, 2))
scatterplotMatrix(~ G557 + G591 + G226 + G718 | type , by.groups =
TRUE,diagonal = 'density', smooth = FALSE, col = c(1, 2),
legend.plot = FALSE, las = 1, reg.line = lm, pch = c(16, 4), data =
genes2)

# los valores t de los genes tt <- tvalues[o[1:4]] names(tt) <-
colnames(genes2)[-2] tt # localizando y organizando los mejores
(genes) selected <- order(abs(tvalues), decreasing = TRUE)[1:30]
newdata <- d[selected,]
```

```

# función para intervalos CV # --- k1 es numero de mejores genes #
k2 es la porción de los datos para dejar por fuera svmk <-
function(k1, k2, kernel){ # selecting the k most DE genes -- DE =
differentially expressed selected <- order(abs(tvalues), decreasing
= TRUE)[1:k1] newdata <- data.frame(t(d[selected,]), cl =
factor(type)) # muestras k <- sample(35) # numero total de
pacientes take <- list(1:3, 4:7, 8:11, 12:15, 16:18, 19:21,
22:25, 26:28, 29:32, 33:35) # grupos balanceados
ks <- lapply(take, function(x) k[x]) # datos de entrenamiento y
modelo SVM lists <- lapply(ks, function(x) newdata[x,]) clas <-
do.call(rbind, lists[-k2]) x <- clas[,-ncol(clas)] y <- clas[,'cl']
model <- svm(x, y, kernel = kernel) # pruebas para datos de
entrenamiento pred <- fitted(model) ta <- table(pred, y)
1-sum(diag(ta))/sum(ta) # clasificación errónea }

# calcular MCR para RL mcrRL <- function(model, cutoff = 0.5, y =
y){
  pred <- predict(model, type = "response")
  yp <- factor(ifelse(pred > cutoff, 1, 0), levels = 0:1)
  ta <- table(yp, y)
  1-sum(diag(ta))/sum(ta)
}

# función para intervalos CV # --- k1 is the number of top genes #
k2 is the portion of the data to leave out RL <- function(k1, k2){ #
seleccionar los k mejores DE genes -- DE = differentially
expressed selected <- order(abs(tvalues), decreasing = TRUE)[1:k1]
newdata <- data.frame(t(d[selected,]), cl = factor(type)) # muestras
k <- sample(35) # total number of patients take <- list(1:3, 4:7,
8:11, 12:15, 16:18, 19:21,
22:25, 26:28, 29:32, 33:35) # grupos balanceados
ks <- lapply(take, function(x) k[x]) # datos de entrenamiento y
modelo RL lists <- lapply(ks, function(x) newdata[x,]) clas <-
do.call(rbind, lists[-k2]) x <- clas[,-ncol(clas)] y <- clas[,'cl']

```

```

y <- as.numeric(y)-1 model <- glm(y ~ ., data = x, family =
binomial)

# pruebas con datos de entrenamiento mcrRL(model, y = y) }

# clasificación errónea -- todos los escenarios juntos

res1 <- sapply(1:30, function(gene)
  sapply(1:10, function(x) svmk(k1 = gene, k2 = x, kernel = 'linear'))))

res2 <- sapply(1:30, function(gene)
  sapply(1:10, function(x) svmk(k1 = gene, k2 = x, kernel = 'polynomial'))))

res3 <- sapply(1:30, function(gene)
  sapply(1:10, function(x) svmk(k1 = gene, k2 = x, kernel = 'radial'))))

res4 <- sapply(1:30, function(gene)
  sapply(1:10, function(x) svmk(k1 = gene, k2 = x, kernel = 'sigmoid'))))

res5 <- sapply(1:30, function(gene)
  sapply(1:10, function(x) RL(k1 = gene, k2 = x)))

r1 <- colMeans(res1) r2 <- colMeans(res2) r3 <- colMeans(res3) r4 <-
colMeans(res4) r5 <- colMeans(res5) out <- cbind(r5, r1, r2, r3, r4)

# ----- # #
usando los 30 mejores genes, uno a uno # #
----- # d1 <-
data.frame(y = factor(as.numeric(factor(type))-1), t(newdata))

svm.foo <- function(many, kernel = 'linear', data = d1){
  d1 <- data[, c(1, 2:(many + 1))]
  fit <- svm(y ~ ., type = "C-classification", kernel = kernel, data = d1)
  pred <- fitted(fit)

```

```
    ta <- table(d1$y, pred)
    1-sum(diag(ta))/sum(ta)
  }

RL.foo <- function(many, data = d1){
  d1 <- data[, c(1, 2:(many + 1))]
  fit <- glm(y ~ ., family = binomial, data = d1)
  pred <- fitted(fit)
  ta <- table(d1$y, ifelse(pred < 0.5, 0, 1))
  1-sum(diag(ta))/sum(ta)
}

# RL unicamente k <- 30 # number of variables RL <- sapply(1:k,
function(g) RL.foo(g)) RL <- data.frame(vars = 1:k, mcr = RL, method
= 'RL')

# todos los SMVs cases <- expand.grid(vars = 1:k, method =
c('linear', 'polynomial', 'radial', 'sigmoid')) cases$mcr <-
apply(cases, 1, function(x) svm.foo(as.numeric(x[1]), x[2], data =
d1)) (cases <- rbind(RL, cases))
```

C. Anexo: Resultados Multivariados Adicionales

En este capítulo se presentan resultados adicionales de la distribución normal multivariada con $p = 10, 20, 50$ que no fueron presentados sección 4.3, por que no su comportamiento no evidenciaba cambios significativos con resultados para $p = 2, 200$.

C.1. Normal Multivariada ($p = 10$)

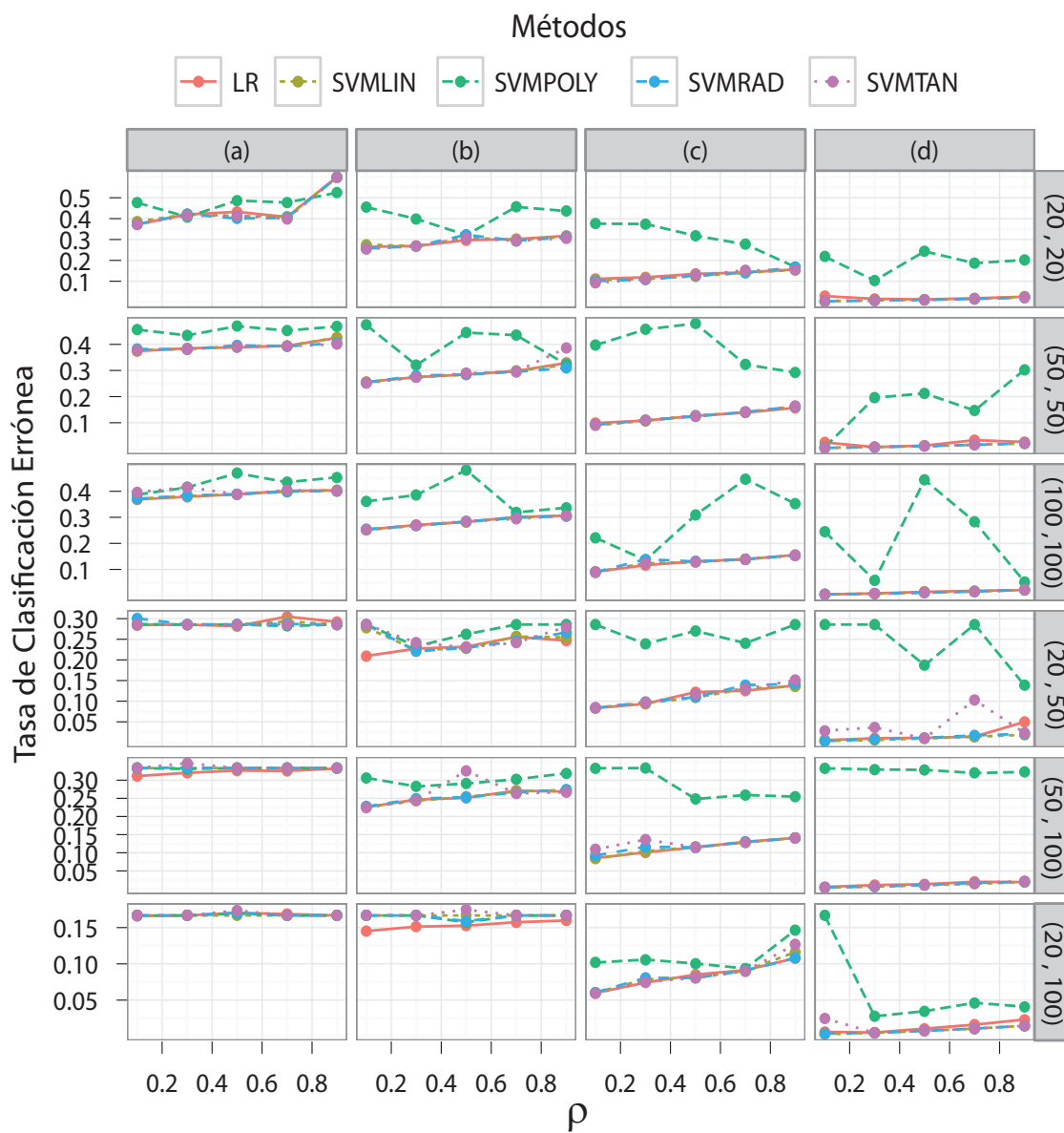


Figura C.1.: Distribución Normal Multivariada ($p = 10$) $\Sigma_1 = \Sigma_2$

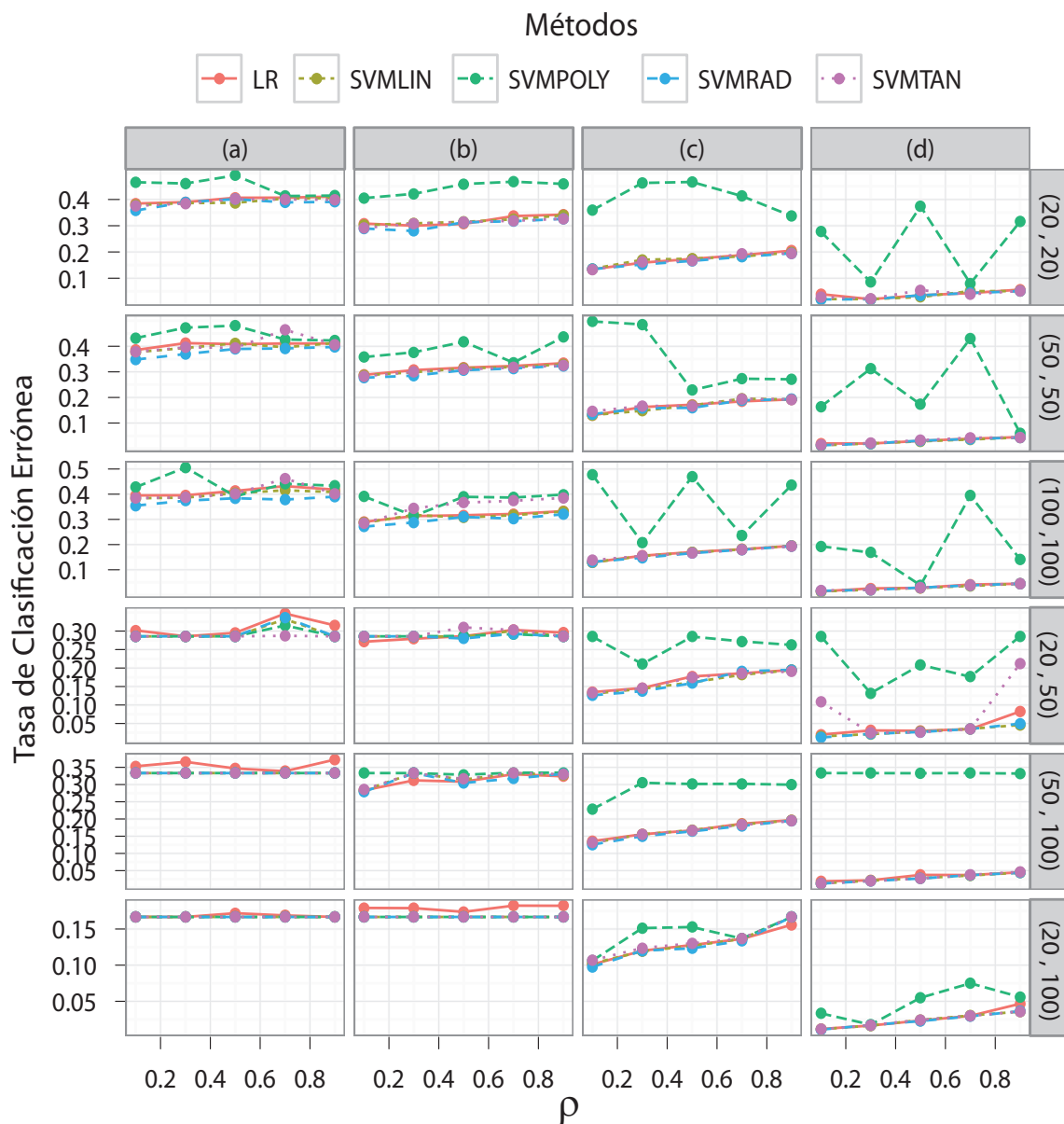


Figura C.2.: Distribución Normal Multivariada ($p = 10$) $\Sigma_1 = 2\Sigma_2$

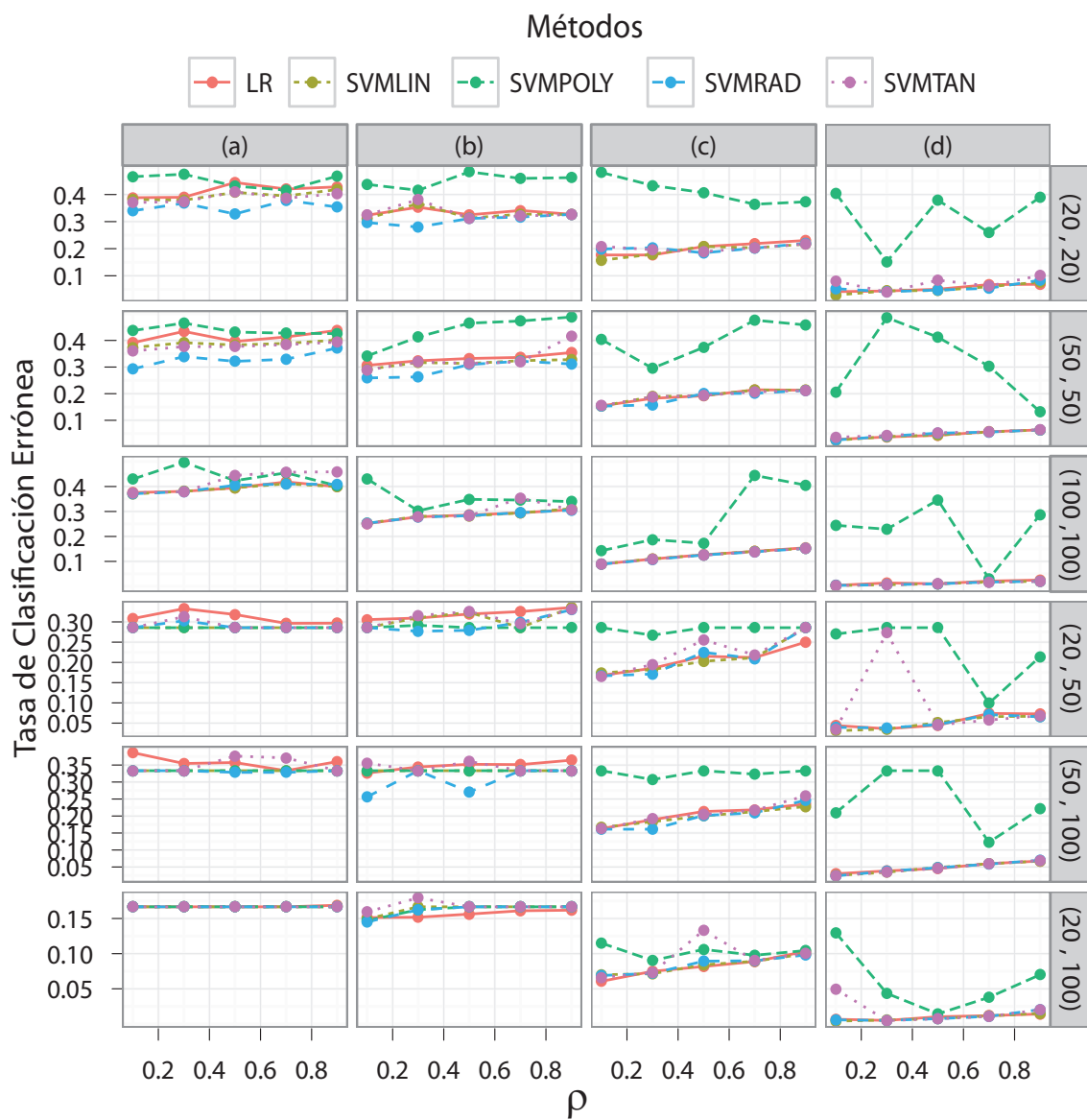


Figura C.3.: Distribución Normal Multivariada ($p = 10$) $\Sigma_1 = 3\Sigma_2$

C.2. Normal Multivariada ($p = 20$)

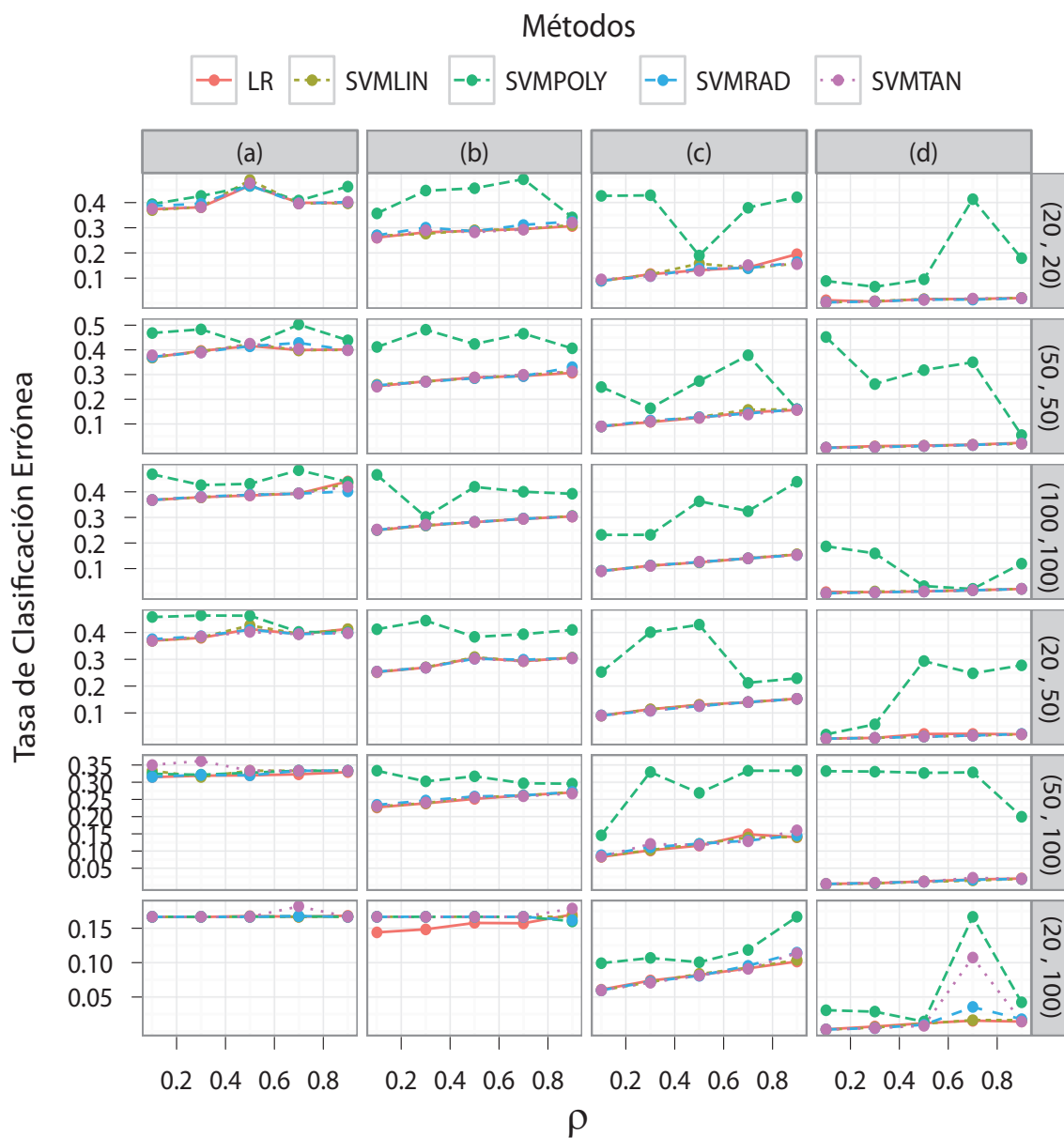


Figura C.4.: Distribución Normal Multivariada ($p = 20$) $\Sigma_1 = \Sigma_2$

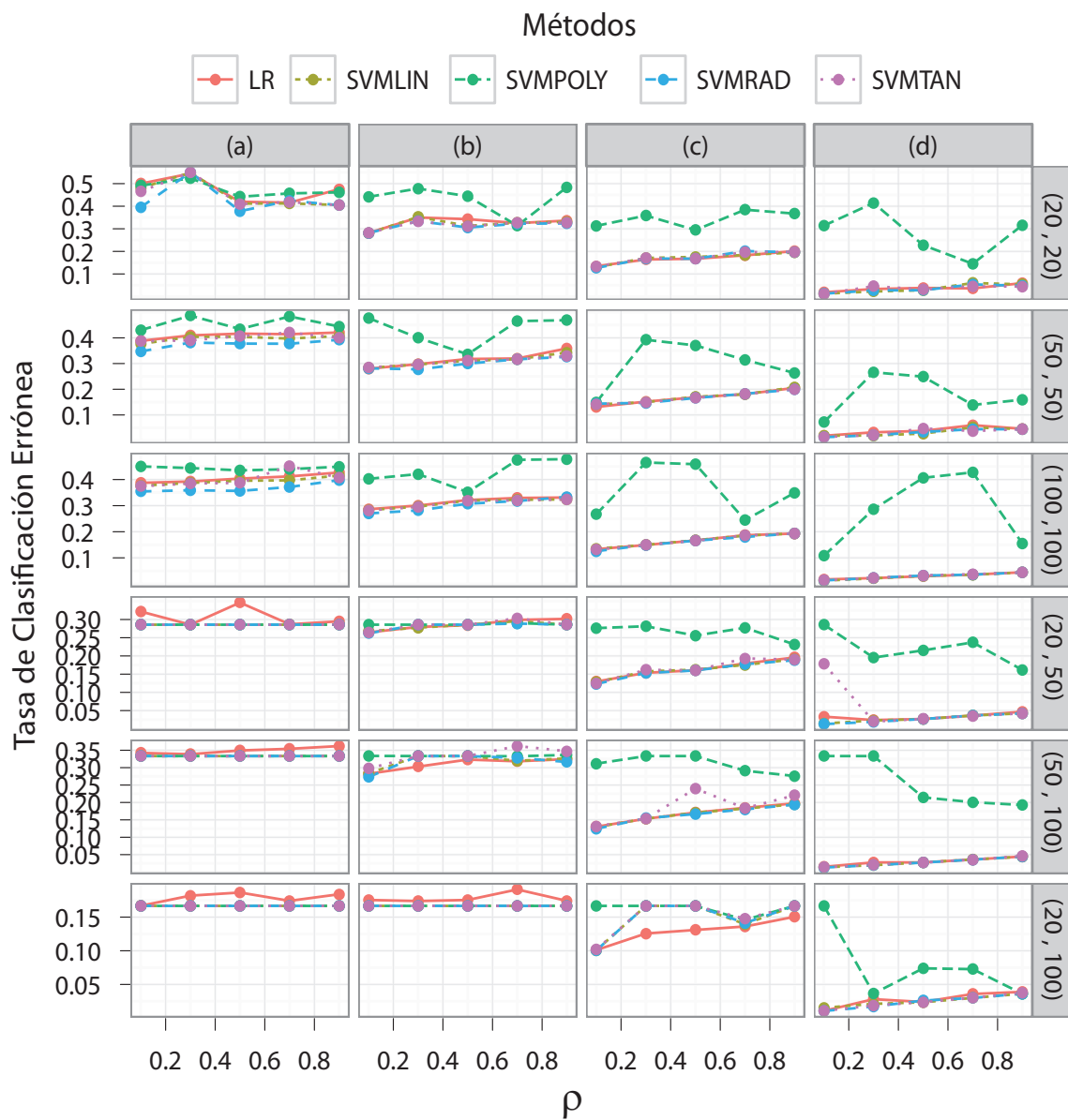


Figura C.5.: Distribución Normal Multivariada ($p = 20$) $\Sigma_1 = 2\Sigma_2$

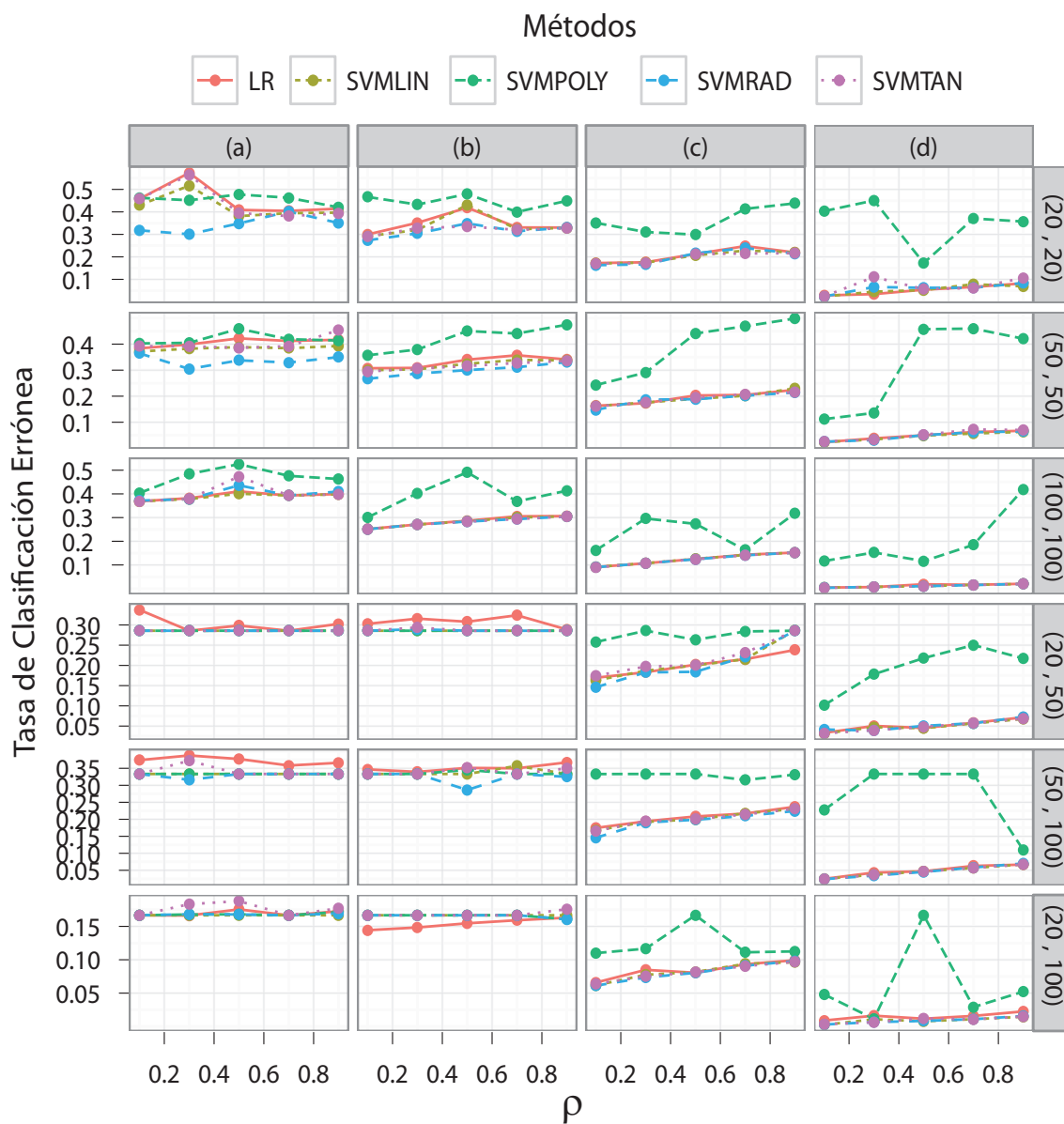


Figura C.6.: Distribución Normal Multivariada ($p = 20$) $\Sigma_1 = 3\Sigma_2$

C.3. Normal Multivariada ($p = 50$)

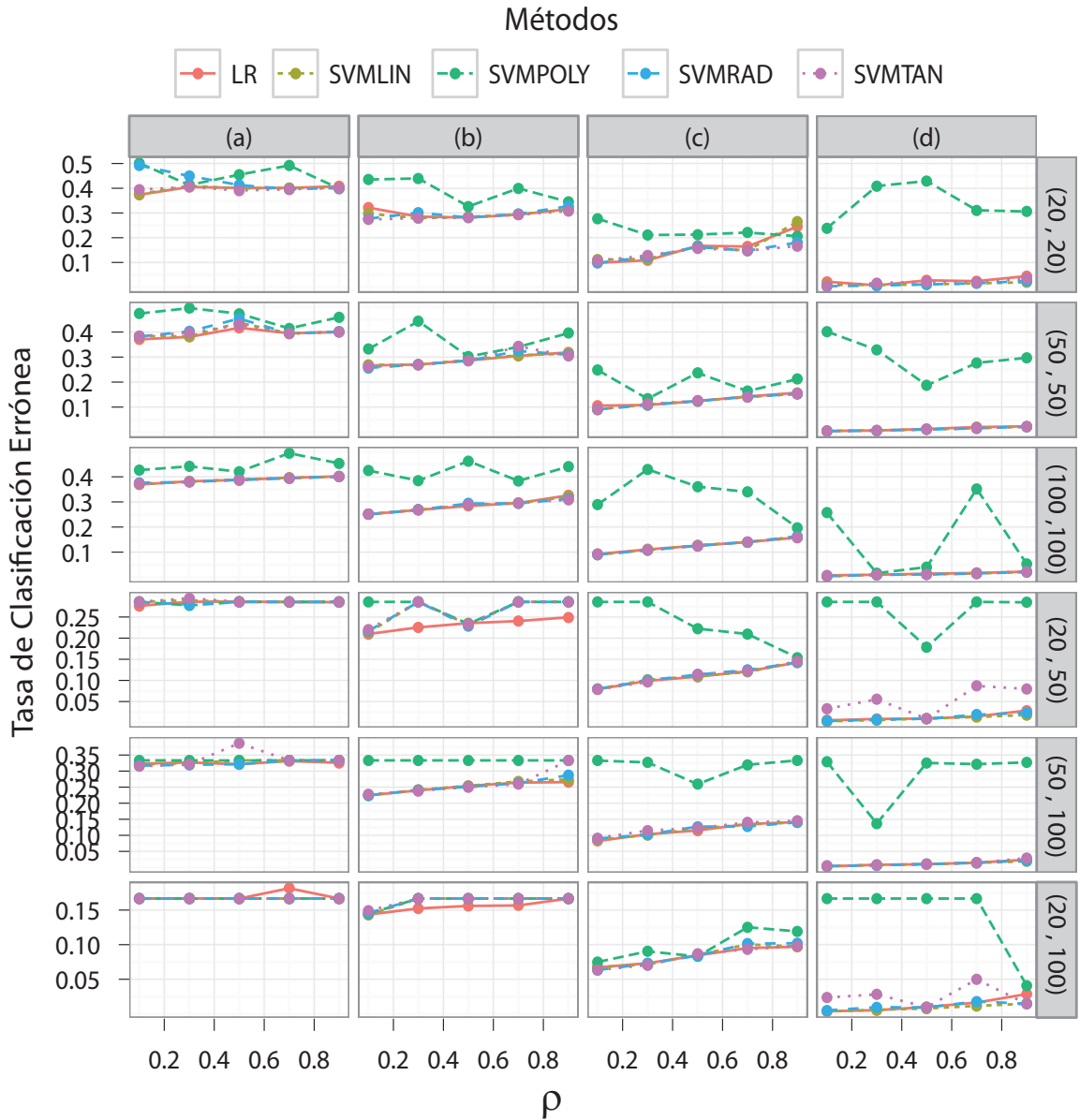


Figura C.7.: Distribución Normal Multivariada ($p = 50$) $\Sigma_1 = \Sigma_2$

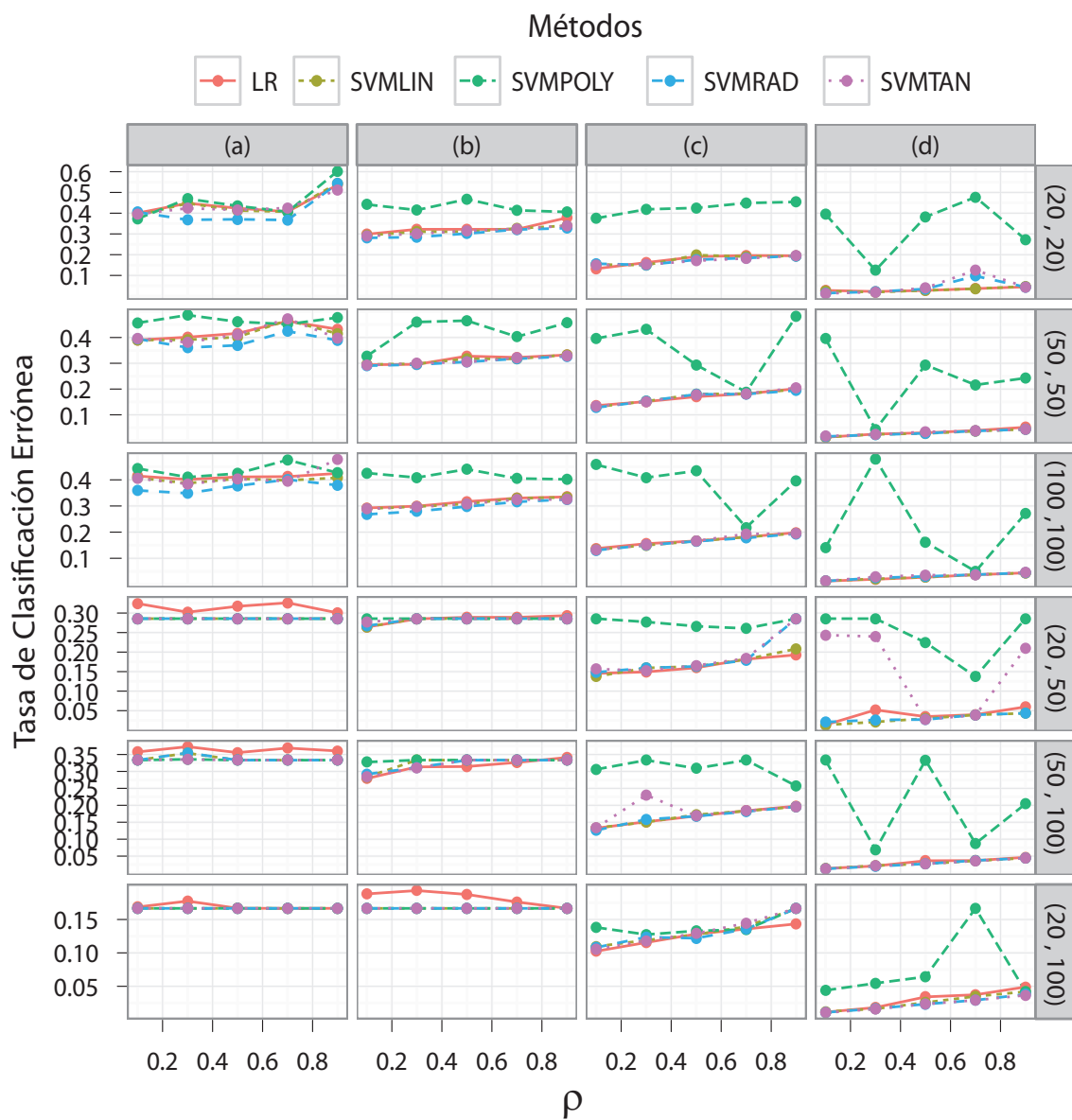


Figura C.8.: Distribución Normal Multivariada ($p = 50$) $\Sigma_1 = 2\Sigma_2$

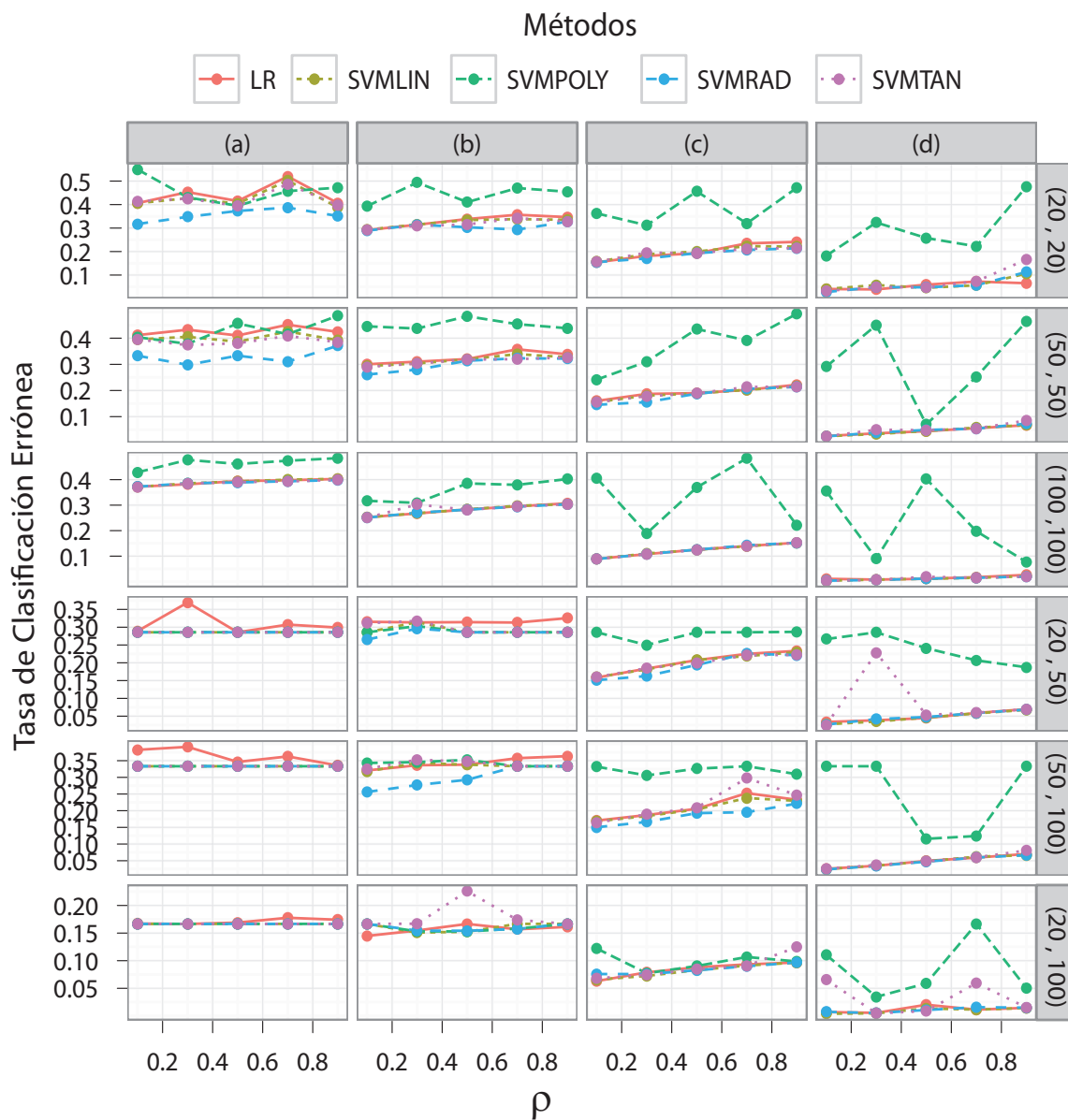


Figura C.9.: Distribución Normal Multivariada ($p = 50$) $\Sigma_1 = 3\Sigma_2$

Bibliografía

- Aizerman, A., Braverman, E. & Rozonoer, L. (1964), 'Theoretical foundations of the potential function method in pattern recognition learning', *Automat. Remote Control* **25**, 821–837.
- ALZ (2011), 'The alzheimer's association', (<http://www.alz.org>). [Fecha de acceso: 5 de Octubre, 2011].
- Anderson, T. (1984), *An introduction to Multivariate Statistical Analysis*, Jhon Wiley & Sons, New York.
- Cornfield, J. (1962), 'Joint dependence of the risk of coronary heart disease on serum cholesterol and systolic blood pressure: A discriminant function analysis', *Proceedings of the Federal American Society of Experimental Biology* **21**, 58–61.
- Correa, J. (2010), 'Diagnósticos de regresión usando la fdr (tasa de descubrimientos falsos)', *Comunicaciones en Estadística* **3**(2), 109–118.
- Cortes, C. & Vapnik, V. (1995), 'Support-vector networks', *Machine Learning* **20**(3), 273–297.
- Cover, T. M. (1965), 'Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition', *IEEE Transactions on Electronic Computers* **14**, 326–334.
- Cox, D. (1966), *Some Procedures Associated with the Logistic Qualitative Response Curve*, Jhon Wiley & Sons, New York.

- Crisler, S., Morrissey, M., Anch, M. & Barnett, D. (2008), ‘Sleep-stage scoring in the rat using a support vector machine’, *Journal of Neuroscience Methods* **168**, 524–534.
- Day, N. & Kerridge, D. (1967), ‘A general maximum likelihood discriminant’, *Biometrics* **23**, 313–323.
- Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D., & Weingessel, A. (2011), *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. R package version 1.5-27.
*<http://CRAN.R-project.org/package=e1071>
- Dubey, A. & Realff, M. (2004), ‘Support vector machines for learning to identify the critical positions of a protein’, *Journal of Theoretical Biology* **243**, 351–361.
- Fisher, R. (1936), ‘The use of multiple measurements in taxonomic problems’, *Annual Eugenics* **7**, 179–188.
- Gentleman, R., Carey, V., Huber, W. & Hahne, F. (2011), *genefilter: genefilter: methods for filtering genes from microarray experiments*. R package version 1.34.0.
- GSE26927 (2011), ‘National center for biotechnology information’, <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26927#>. [Fecha de acceso: 27 de Septiembre, 2011].
- Hernández, F. & Correa, J. (2009), ‘Comparación entre tres técnicas de clasificación’, *Revista Colombiana de Estadística* **32**(2), 247–265.
- Hongdong, L., Yizeng, L. & Qingsong, X. (2009), ‘Support vector machines and its applications in chemistry’, *Chemometrics and Intelligent Laboratory Systems* **95**, 188–198.
- Hosmer, D. & Lemeshow, S. (1989), *Applied Logistic Regression*, Jhon Wiley & Sons, New York.

- Houston, E. & Woodruff, D. (1997), *Empirical Bayes Estimates of Parameters from the Logistic Regression Model*, ACT Research Report Series 97-6.
- Karatzoglou, A., Meyer, D. & Hornik, K. (2006), ‘Support vector machines in R’, *Journal of Statistical Software* **15**(8), 267–73.
- Lee, J. B., Park, M. & Song, H. S. (2005), ‘An extensive comparison of recent classification tools applied to microarray data’, *Computational Statistics & Data Analysis* **48**, 869–885.
- Lu, C., Van Gestel, T., J.A., S., Van Huffel, S., Vergote, I. & Timmerman, D. (2003), ‘Preoperative prediction of malignancy of ovarian tumors using least squares support vector machines’, *Artificial Intelligence in Medicine* **28**(3), 281–306.
- Moguerza, J. & Muñoz, A. (2006), ‘Vector machines with applications’, *Statistical Science* **21**(3), 322–336.
- Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D. & Groop, L. C. (2003), ‘Pgc-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes’, *Nat. Genet.* **34**(3), 267–73.
- NCBI (2011), ‘National center for biotechnology information’, <http://www.ncbi.nlm.nih.gov>. [Fecha de acceso: 20 de Septiembre, 2011].
- Nguyen, D. V., Bulak Apart, A., Wang, N. & Carrol, R. J. (2002), ‘Dna microarray experiments:biological and technological aspects’, *Biometrics* **58**, 701–717.
- Piegorsch, W. & Casella, G. (1996), ‘Empirical bayes estimation for logistic regression and extended parametric regression’, *Journal of Agricultural, Biological, and Environmental Statistics* **1**(2), 231–249.

- R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
*<http://www.R-project.org/>
- Shou, T., Hsiao, Y. & Huang, Y. (2009), ‘Comparative analysis of logistic regression, support vector machine and artificial neural network for the differential diagnosis of benign and malignant solid breast tumors by the use of three-dimensional power doppler’, *Korean J Radiol* **10**, 464–471.
- Tibshirani, R. & Friedman, J. (2008), *The Elements of Statistical Learning*, Springer, California.
- Tripathi, S., Srinivas, V. & Nanjundiah, R. (2006), ‘Downscaling of precipitation for climate change scenarios: A supportvectormachine approach’, *Journal of Hydrology* **330**, 621–640.
- Twyman, R. (2003), ‘Human genome website’, http://genome.wellcome.ac.uk/doc_WTD020757.html. [Fecha de acceso: 24 de Marzo, 2012].
- Vapnik, V. & Chervonenkis, A. (1969), ‘Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies’, *Automat. Remote Control* **25**, 103–109.
- Vélez, J. (2008), Comparación de 4 procedimientos fdr para la selección de parámetros en regresión poisson, Tesis de Maestría, Escuela de Estadística, Universidad Nacional de Colombia, Sede Medellín.
- Verplancke, T., Van Looy, S., Benoit, D., Vansteelandt, S., Depuydt, P., De Turck, F. & Decruyenaere, J. (2008), ‘Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies’, *BMC Med. Inform. Decis. Mak.* **8**, 56–64.
- Westreich, D., Lessler, J. & Jonsson, M. (2010), ‘Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-

classifiers as alternatives to logistic regression', *Journal of Clinical Epidemiology* **63**, 826–833.

Whittemore, A. (1995), 'Logistic regression of family data from case-control studies', *Biometrika* **82**(1), 57–67.

Whittemore, A. (2004), 'Estimating genetic association parameters from family data', *Biometrika* **91**(1), 219–225.