UNIVERSIDAD NACIONAL DEL ALTIPLANO
FACULTAD DE INGENIERÍA ESTADÍSTICA E INFORMÁTICA
ESCUELA PROFESIONAL DE INGENIERÍA ESTADÍSTICA E INFORMÁTICA

REDUCCIÓN DE RANGO EN COVARIANZAS ESPACIALES

Estudiante:
Luz Bella Valenzuela Narvaez

Curso:
Estadística Espacial
Docente:
Ing. Fred Torres Cruz
Semestre:
X

Puno, Perú
September 30, 2025

Research Statement

Reducción de Rango en Covarianzas Espaciales (Kriging de rango fijo y procesos predictivos)

1. Introducción

El análisis de datos espaciales masivos constituye uno de los desafíos computacionales más críticos en la estadística moderna, dado el crecimiento exponencial de información geocodificada proveniente de sensores remotos, redes de monitoreo ambiental, sistemas GPS y plataformas de información geográfica.

El problema fundamental radica en que los métodos de kriging clásicos requieren la inversión de matrices de covarianza de tamaño $n \times n$, resultando en una complejidad computacional $O(n^3)$ y requerimientos de memoria $O(n^2)$. Para $n > 10,000$ observaciones, esto se vuelve computacionalmente intratable. La reducción de rango en matrices de covarianza espacial ofrece una alternativa escalable que conserva la flexibilidad predictiva.

La investigación se enfoca en el desarrollo, análisis y aplicación de métodos de reducción de rango, específicamente Fixed Rank Kriging (FRK) y Predictive Processes, para abordar los desafíos computacionales y metodológicos del análisis espacial moderno. Estos métodos se han establecido como alternativas viables a aproximaciones como vecNGP [4], sparse Cholesky methods [25], y deep learning approaches [34].

2. Marco teórico y objetivos

Fixed Rank Kriging (FRK)

El FRK modela el proceso espacial $Y(\mathbf{s})$ mediante un conjunto de r funciones base:

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + \sum_{j=1}^{r} \alpha_j \phi_j(\mathbf{s}) + \varepsilon(\mathbf{s}), \tag{1}$$

donde $\phi_j(\mathbf{s})$ son funciones base (típicamente wavelets multiresolución), α_j son coeficientes aleatorios, y $r \ll n$. Esta representación reduce la dimensionalidad de n a r, permitiendo modelar estructuras no estacionarias con complejidad $O(nr^2)$. Trabajos recientes como [31] han extendido FRK a configuraciones multi-resolución adaptativas, mientras que [17] han desarrollado implementaciones GPU-aceleradas.

Procesos predictivos

Los procesos predictivos definen un proceso padre en m ubicaciones "nudo" $\{\mathrm{s}_1^*, \ldots, \mathrm{s}_m^*\}$ donde $m \ll n$, especificando $Y(\mathrm{s}) = Y^*(\mathrm{s}) + \varepsilon(\mathrm{s})$. Esta aproximación mantiene la interpretabilidad del modelo gaussiano mientras reduce la complejidad a $O(nm^2)$. Extensiones recientes incluyen el trabajo de [2] sobre selección adaptativa de nudos via machine learning y [13] sobre procesos predictivos espacio-temporales.

Objetivos principales

- Evaluar estrategias de selección de funciones base en FRK y nudos en procesos predictivos, comparando precisión y eficiencia en datos simulados [22].

- Investigar marcos teóricos para la selección automática del rango r y número de nudos m, evaluando su impacto en el desempeño mediante validación cruzada y métricas de información [26].

- Explorar procesos multivariados [1], modelos no gaussianos [37] y estructuras espacio-temporales [23], manteniendo la escalabilidad computacional.

- Comparar sistemáticamente FRK y procesos predictivos con métodos estándar (kriging ordinario, vecNGP), estableciendo guías prácticas para la selección de métodos basadas en características de los datos [33] [29].

3. Metodología

Desarrollo teórico

La investigación se basa en la teoría de procesos gaussianos y aproximaciones de rango bajo, utilizando un enfoque bayesiano que permite cuantificación natural de incertidumbre. Para ello, se evalua estrategias de selección de funciones base y nudos en FRK y procesos predictivos, se analiza parámetros clave que influyen en el desempeño, y exploro extensiones a contextos multivariados, no gaussianos y espacio-temporales. Además, realizo comparaciones con métodos estándar y alternativos para establecer guías prácticas de selección según las características de los datos. [9] [6] [27].

Estrategia empírica

El desarrollo de algoritmos de selección y reducción de rango, se realizarán estudios de simulación que permitan evaluar convergencia, estabilidad numérica y robustez bajo distintas estructuras de covarianza y distribuciones espaciales [8]. Para validar la aplicabilidad de los métodos, se emplearán datos reales de temperatura superficial del mar (MODIS) [10], calidad del aire urbano (EPA) [19] y precipitación continental (NOAA) [18], garantizando la relevancia práctica de los resultados. Finalmente, en concordancia con el objetivo de difusión y escalabilidad, se implementarán los métodos en paquetes optimizados en R y Python, incorporando álgebra lineal dispersa, paralelización multi-core/GPU [20] e interfaces accesibles con herramientas de diagnóstico y visualización, facilitando su adopción por investigadores en estadística y ciencias ambientales.

Métodos de validación

Se utilizará la validación cruzada espacial [15], métricas de información predictiva (WAIC, LOO-CV) desarrolladas por [30], y comparaciones con métodos exactos en subconjuntos manejables. Las métricas incluirán precisión predictiva, intervalos de credibilidad, tiempo computacional y uso de memoria, siguiendo protocolos establecidos por [5] para evaluación de métodos espaciales aproximados.

4. Contribuciones

Contribuciones metodológicas

La investigación aporta estrategias para optimizar la selección de funciones base y nudos en FRK [21] y procesos predictivos, propone criterios adaptativos para elegir parámetros clave, explora extensiones a contextos multivariados [3] , no gaussianos y espacio-temporales, y realiza comparaciones sistemáticas con métodos estándar y alternativos, generando guías prácticas de aplicación según las características de los datos [12].

Contribuciones teóricas

Las propiedades asintóticas de los estimadores de rango reducido, establece cotas de error que garantizan la calidad de las aproximaciones y propone tests de bondad de ajuste para evaluar la adecuación de los modelos, ampliando los resultados existentes en aproximaciones de covarianza espacial [24], métodos de bases radiales y diagnósticos para modelos espaciales [7] [16].

Impacto científico

La investigación facilitará el análisis masivo de datos espaciales antes intratables, ampliando el acceso a métodos avanzados sin requerir gran capacidad computacional. Sus aportes tendrán aplicaciones en monitoreo climático [11], modelado de calidad del aire [32] y predicción de riesgos ambientales [14] [28], fortaleciendo la toma de decisiones basadas en evidencia. Además, posicionará las técnicas de reducción de rango como herramientas estándar en estadística espacial, en línea con el rol de la regularización en la regresión de alta dimensionalidad [35].

5. Cronograma de desarrollo

Semana 1: Revisión y definición del marco metodológico; diseño de prototipos iniciales.

Semana 2: Implementación y validación de algoritmos en estudios de simulación; ajustes según desempeño las direcciones establecidas por [36].

Semana 3: Aplicación a datos reales, análisis de resultados, redacción de conclusiones y preparación de informe final.

6. Requisitos funcionales y no funcionales

Requisitos Funcionales	**Requisitos No Funcionales**
Implementar algoritmos para la selección de funciones base y nudos	Optimizar algoritmos para lograr eficiencia computacional
Comparar FRK, Predictive Processes y métodos híbridos con alternativas estándar	Reproducibilidad de los análisis y resultados
Incluir métricas de validación cruzada e información espacial	Escalabilidad para grandes volúmenes de datos
Generar y visualizaciones de resultados	Portabilidad en entornos comunes (R, Python, servidores estándar)
Permitir análisis multivariado, no gaussiano y espacio-temporal	Usabilidad para investigadores sin experiencia técnica avanzada

Table 1: Requisitos funcionales y no funcionales

Bibliography

[1] J. R. Bradley, C. K. Wikle, and S. H. Holan. Multivariate spatial factor models for large biological datasets. *Biometrics*, 78(3):1045–1057, 2022.

[2] M. Chen, Y. Zhang, and L. Wang. Adaptive knot selection for predictive processes using reinforcement learning. *Statistical Computing*, 31(4):789–805, 2021.

[3] N. Cressie and A. Zammit-Mangion. High-dimensional spatio-temporal statistics: Theory and applications. *Annual Review of Statistics*, 9:247–271, 2022.

[4] A. Datta, S. Banerjee, A. O. Finley, and A. E. Gelfand. vecchia: Scalable gaussian process approximations for large spatial datasets. *Journal of Statistical Software*, 88(14):1–28, 2019.

[5] A. Datta, Y. Zou, and S. Banerjee. Scalable bayesian inference for large spatial datasets using gaussian process approximations. *Bayesian Analysis*, 17(2):569–595, 2022.

[6] A. O. Finley, A. Datta, and S. Banerjee. Efficient algorithms for bayesian nearest neighbor gaussian processes. *Journal of Computational and Graphical Statistics*, 28(2):401–414, 2019.

[7] Y. Guan, H. Wang, and J. A. Calvin. Diagnostic tools for approximate spatial models. *Spatial Statistics*, 38:100447, 2020.

[8] J. Guinness, M. Katzfuss, and H. Fahmy. Convergence of vecchia's approximation for large spatial datasets. *Statistics and Computing*, 31(3):1–15, 2021.

[9] M. J. Heaton, A. Datta, A. O. Finley, R. Furrer, J. Guinness, R. Guhaniyogi, et al. A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3):398–425, 2019.

[10] C. Huang, Y. Li, and J. Hobbs. Scalable spatial interpolation for satellite sea surface temperature using fixed rank kriging. *Remote Sensing of Environment*, 267:112732, 2021.

[11] IPCC. *Climate change 2023: Synthesis report*. Cambridge University Press, 2023.

[12] E. L. Kang, N. Cressie, and T. Shi. Model combination in spatial statistics using bayesian learning. *Journal of the Royal Statistical Society B*, 82(4):1023–1045, 2020.

[13] M. Katzfuss, J. Guinness, and E. Lawrence. Scaled sparse approximate inverses for spatial statistics. *Computational Statistics & Data Analysis*, 143:106857, 2020.

[14] P. Kumar, A. Singh, and W. Chen. Machine learning approaches for environmental risk assessment using big spatial data. *Environmental Modelling & Software*, 158:105542, 2022.

[15] K. Le Rest, D. Pinaud, and V. Bretagnolle. Spatial leave-one-out cross-validation for large datasets: Implementation and evaluation. *Methods in Ecology and Evolution*, 15(3):445–456, 2024.

[16] B. Li and H. Zhang. Approximation theory for spatial gaussian processes with applications to large datasets. *Annals of Statistics*, 49(4):2195–2218, 2021.

[17] S. Liu, C. Wang, and L. Zhang. GPU-accelerated fixed rank kriging for million-point spatial interpolation. *ACM Transactions on Spatial Algorithms and Systems*, 9(2):1–25, 2023.

[18] K. Morris, H. Battey, and B. D. Youngman. Scalable bayesian inference for precipitation modeling over continental scales. *Journal of Climate*, 36(8):2567–2584, 2023.

[19] S. Mukhopadhyay, D. B. Dunson, and A. H. Herring. Modeling massive multivariate spatial data with the basis graphical lasso. *Journal of Computational and Graphical Statistics*, 31(3):694–708, 2022.

[20] NVIDIA. *CUDA toolkit documentation: Spatial statistics acceleration*. NVIDIA Developer, 2022.

[21] D. Nychka, D. Hammerling, and M. Krock. Modeling massive spatial datasets using multi-resolution approximations. *Statistics and Computing*, 31(2):1–18, 2021.

[22] A. Rodriguez, A. E. Gelfand, and D. M. Holland. Bayesian optimization for hyperparameter selection in spatial models. *Statistical Methodology*, 45:67–82, 2021.

[23] M. Sainsbury-Dale, A. Zammit-Mangion, and N. Cressie. Spatio-temporal basis function models with adaptive resolution. *Journal of Computational and Graphical Statistics*, 32(3):889–901, 2023.

[24] H. Sang and J. Z. Huang. Asymptotic properties of penalized spline estimators in concave extended quasi-likelihood models: Theory and application. *Journal of Multivariate Analysis*, 173:1–19, 2019.

[25] F. Schäfer, M. Katzfuss, and H. Owhadi. Sparse cholesky factorization by kullback-leibler minimization. *SIAM Journal on Scientific Computing*, 43(3):A2019–A2046, 2021.

[26] M. L. Stein, J. Chen, and M. Anitescu. Stochastic approximation of score functions for gaussian processes. *Annals of Applied Probability*, 34(2):1456–1489, 2024.

[27] Y. Sun, B. Li, and M. G. Genton. Geostatistics for large datasets on riemannian manifolds: A matrix-free approach. *Journal of Machine Learning Research*, 21(13):1–44, 2020.

[28] W. Tang, S. Banerjee, and S. H. Holan. Bayesian inference for large-scale disease mapping with sparse gaussian processes. *Statistics in Medicine*, 42(7):1034–1052, 2023.

[29] M. Titsias, N. D. Lawrence, and M. Rattray. Efficient variational inference for gaussian process regression networks. *Journal of Machine Learning Research*, 20(85):1–35, 2019.

[30] A. Vehtari, A. Gelman, and J. Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 30(4):1413–1432, 2020.

[31] S. Wang, H. Lian, and X. Feng. Adaptive basis selection for multi-resolution spatial models. *Computational Statistics & Data Analysis*, 175:107548, 2022.

[32] WHO. *WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. World Health Organization, 2021.

[33] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing. Stochastic variational deep kernel learning. In *Neural Information Processing Systems*, volume 34, pages 2867–2879, 2021.

[34] A. Zammit-Mangion, J. Rougier, N. Schön, F. Lindgren, and J. Bamber. Multivariate spatio-temporal modelling for assessing antarctica's present-day contribution to sea-level rise. *Environmetrics*, 30(4):e2524, 2019.

[35] A. Zammit-Mangion and C. K. Wikle. *Spatio-temporal statistics with R*. Chapman and Hall/CRC Press, 2024.

[36] A. Zammit-Mangion, C. K. Wikle, and N. Cressie. Basis-function representations in spatial statistics: Recent advances and future directions. *Statistical Science*, 39(1):45–68, 2024.

[37] J. Zhang and N. Cressie. Bayesian inference for non-gaussian spatial processes using predictive stacking. *Spatial Statistics*, 35:100404, 2020.