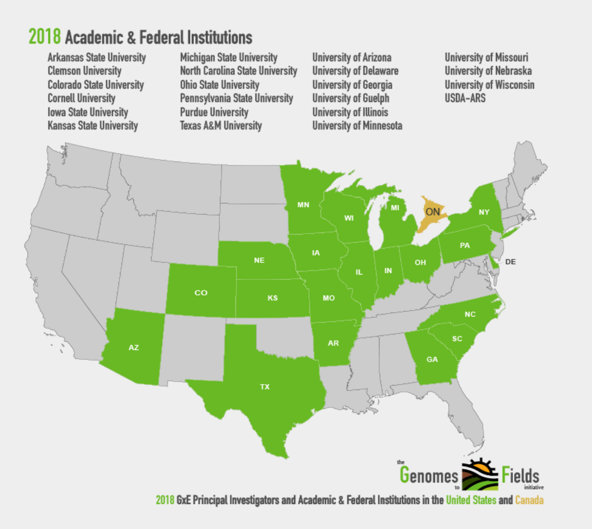
A Hierarchical Spatial Finlay-Wilkinson Model for Analysis of Multi-Environment Field Trials

Xingche Guo, Somak Dutta, Dan Nettleton

Dept. of Statistics, Iowa State University

Second International Workshop on:
Machine Learning for Cyber-Agricultural Systems

The Genomes to Fields (G2F) Initiative



copyright: `https://www.genomes2fields.org`

Multi-Environment Field Trial Analysis for G2F Data

We only focus on 2015 G2F dataset with:

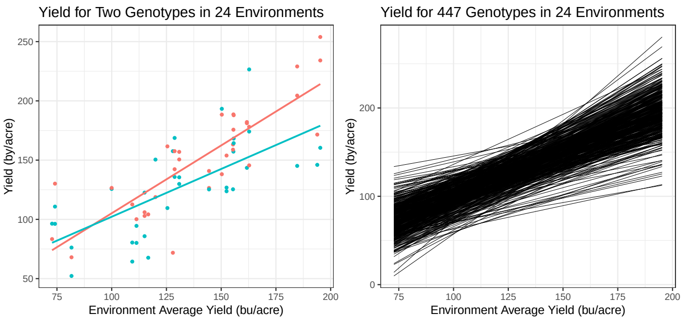
- A subset of 24 environments (field trials).
- Yield recorded for 10,971 field plots with known spatial locations.
- A total of 1,105 hybrid genotypes (varieties).
- SNPs sequence data at $\sim$1M genomic loci are available.
- Time-indexed measurements for weather variables (temperature, rainfall amount, solar radiation, etc), and several soil variables (pH value, soil organic matter, etc).

Finlay-Wilkinson (FW) Model

- Finlay-Wilkinson (FW) model (Finlay and Wilkinson, 1963):

$$y_{ijk} = \mu + g_i + h_j + b_i h_j + e_{ijk},$$

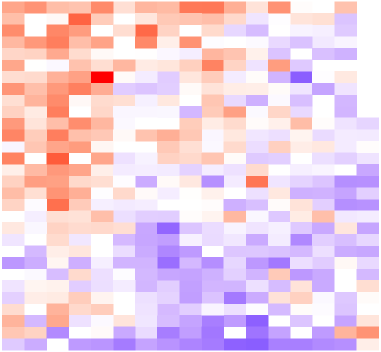
- where μ is the overall mean, g_i is the genotype effect, h_j is the environment effect, $b_i h_j$ is the FW-type multiplicative interaction effect.



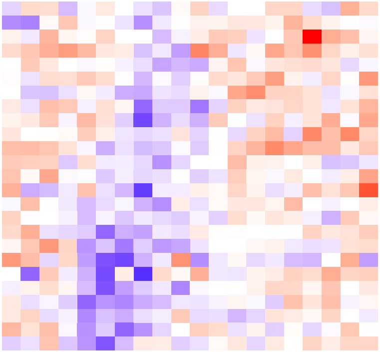
Residuals of FW Model for Two Fields

Problem: the residuals are highly spatially correlated.

MOH1



WIH2



Hierarchical Spatial Finlay-Wilkinson (SFW) Model

- Data model:

$$[y_{ijk}|\mu, \mathbf{g}, \mathbf{b}, \mathbf{h}, \phi] \overset{indep}{\sim} \mathcal{N}(\mu + g_i + h_j + b_i h_j + \phi_{ijk}, \sigma_e^2),$$

- Prior distributions for genotype, slope, and field effects:

$$[\mathbf{g}] \sim \mathcal{N}(\mathbf{0}, \mathbf{A}\sigma_g^2); \quad [\mathbf{b}] \sim \mathcal{N}(\mathbf{0}, \mathbf{A}\sigma_b^2);$$

$$[\mathbf{h}|\gamma] \sim \mathrm{N}(\gamma_1 \mathbf{Z}_1 + \cdots + \gamma_l \mathbf{Z}_l + \cdots + \gamma_L \mathbf{Z}_L, \mathbf{I}\sigma_h^2).$$

- **A** is the kinship matrix describing the correlation structure between different hybrid corn varieties (rrBLUP in R, Tassel 5).

- $\mathbf{Z}_l$ is the lth standardized environmental covariate.

Intrinsic Autoregression Model for Spatial Effects

- A popular model for fertility adjustment in agricultural field trials is the first order intrinsic autoregression (Besag and Higdon, 1999; Dutta and Mondal, 2015).
- First order Intrinsic Autoregressive prior:

$$[\boldsymbol{\psi}_j|\theta_j, \sigma_j^2] \propto |\sigma_j^{-2}\mathbf{W}_j|_+^{1/2} \exp\left(-\tfrac{1}{2}\sigma_j^{-2}\boldsymbol{\psi}_j^{\mathrm{T}}\mathbf{W}_j\boldsymbol{\psi}_j\right)$$

  where

$$\boldsymbol{\psi}_j^{\mathrm{T}}\mathbf{W}_j\boldsymbol{\psi}_j = \theta_j \sum\sum(\psi_{u,v} - \psi_{u-1,v})^2 + \bar{\theta}_j \sum\sum(\psi_{u,v} - \psi_{u,v-1})^2$$

- The distribution of $\boldsymbol{\psi}_j$ is invariant to the addition of $c\mathbf{1}$.

Intrinsic Autoregression Model for Spatial Effects

Recall:

- $[y_{ijk}|\mu, \mathbf{g}, \mathbf{b}, \mathbf{h}, \phi] \overset{indep}{\sim} \mathcal{N}(\mu + g_i + h_j + b_i h_j + \phi_{ijk}, \sigma_e^2),$

Intrinsic Autoregression Model for Spatial Effects

Recall:

- $[y_{ijk}|\mu, \mathbf{g}, \mathbf{b}, \mathbf{h}, \phi] \overset{indep}{\sim} \mathcal{N}(\mu + g_i + h_j + b_i h_j + \phi_{ijk}, \sigma_e^2)$,

Problem:

- The mean of intrinsic spatial prior is not well-defined.
- The overall levels of spatial effects are confounded with the environment effects.
- Estimation of $\mathbf{b}$ is biased.
- Hierarchical structure of $\mathbf{h}$ is not applicable.

Intrinsic Autoregression Model for Spatial Effects

Recall:

- $[y_{ijk}|\mu, \mathbf{g}, \mathbf{b}, \mathbf{h}, \phi] \overset{indep}{\sim} \mathcal{N}(\mu + g_i + h_j + b_i h_j + \phi_{ijk}, \sigma_e^2),$

Problem:

- The mean of intrinsic spatial prior is not well-defined.
- The overall levels of spatial effects are confounded with the environment effects.
- Estimation of $\mathbf{b}$ is biased.
- Hierarchical structure of $\mathbf{h}$ is not applicable.

Solution:

- A hard constraint: set the average of the spatial effects to zero.

Projected Intrinsic Autoregression (PIAR) Prior

- The Gaussian projected intrinsic autoregression (PIAR) prior on the $r_j \times c_j$ regular array is defined as:

$$\phi_j = \mathbf{B}_j \varphi_j, \qquad \varphi_j \sim \mathcal{N}(\mathbf{0}, \mathbf{D}_j^{-1}),$$

- A sum-to-zero constrained version of intrinsic autoregression prior.
- $\mathbf{D}_j$ is the $(r_j c_j - 1) \times (r_j c_j - 1)$ diagonal matrix with its diagonal entries to be all the nonzero eigenvalues of $\mathbf{W}_j$.
- $\mathbf{B}_j$ is the $r_j c_j \times (r_j c_j - 1)$ corresponding eigenvector matrix.

Matrix Free Computation

- The covariance matrix of the Gaussian PIAR is a dense singular matrix.
- The computation load for generating ϕ_j from PIAR using knowledge of multivariate statistics is $\mathcal{O}(M_j^3/3)$, where $M_j = r_j c_j$.
- Assume small number of missing plots (denote $m_j := M_j - N_j$ as the number of missing plots, we assume $m_j \ll M_j$).
- Thus matrix-vector multiplications with $\mathbf{B}_j$ and $\mathbf{B}_j^{\mathrm{T}}$ can also be performed using these discrete cosine transformations (DCT).
- The computation load of our proposed algorithm is $\mathcal{O}(M_j + m_j M_j \log M_j + m_j^3/3)$.

Prediction

- Implement posterior predictive distributions.
- Easy to obtain predictive credible intervals.

Prediction

- Implement posterior predictive distributions.
- Easy to obtain predictive credible intervals.

Within-field prediction:

- Important to account for the spatial correlation between plots.
- Kinship information plays a decisive role for an accurate prediction.
- Mainly used for model evaluation.

Prediction

- Implement posterior predictive distributions.
- Easy to obtain predictive credible intervals.

Within-field prediction:

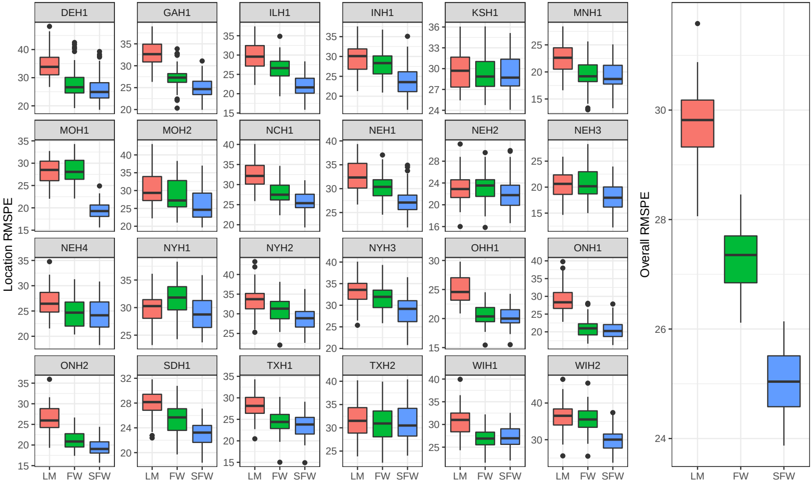
- Important to account for the spatial correlation between plots.
- Kinship information plays a decisive role for an accurate prediction.
- Mainly used for model evaluation.

Predict in new environments:

- By learning how environment effects depend on the weather and soil variables.

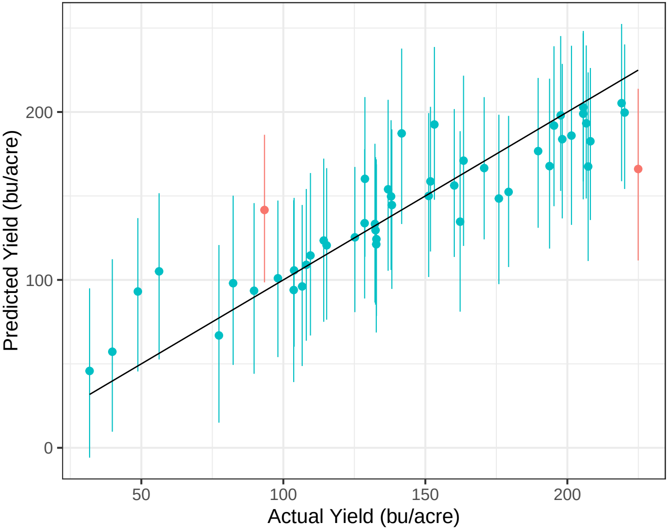
Model Evaluation via Within-Field Prediction

Reduced error for yield prediction.



Prediction Intervals

50 plot yield prediction intervals (95% credible level).



Prediction Intervals Width

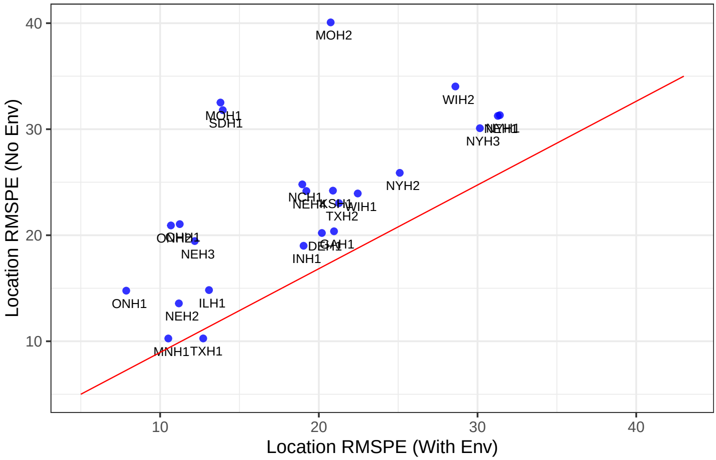
the median credible interval widths of LM, FW, and SFW models at 90% *and* 95% *credible levels are provided.*

	90% CL			95% CL		
	LM	FW	SFW	LM	FW	SFW
Coverage Percentages	90.3%	89.9%	90.1%	95.3%	94.9%	94.5%
Median Interval Widths	98.4	90.3	80.1	117.3	107.5	95.76

SFW model has a more precise interval prediction given that SFW model has the shortest interval widths at the same coverage levels.

Predict in New Environments

Location-wise RMSPEs computed using temperature and rainfall data (x-axis), versus the location-wise RMSPEs computed not using any environment information (y-axis).



Our contribution

- Proposed a unified framework for high-dimensional GxE analysis by integrating genomic, environmental, and within-field spatial information.
- Proposed PIAR prior and its fast computation algorithm in MCMC for multi-environment trials analysis.
- Allow us to predict the yield of a (possibly novel) corn variety in a (possibly new) environment.

What's next

- Allow more complex models (non-linear models, time series models, functional data models, etc) for environmental covariates.
- Formulate better kinship matrix to improve estimation and further accelerate the algorithm.
- Extend to generalized HSFW model to account for discrete value responses.

Selected References

Besag, J. and Higdon, D. (1999). Bayesian analysis of agricultural field experiments. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(4):691–746.

Dutta, S. and Mondal, D. (2015). An h-likelihood method for spatial mixed linear models based on intrinsic auto-regressions. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 77(3):699–726.

Finlay, K. and Wilkinson, G. (1963). The analysis of adaptation in a plant-breeding programme. Australian Journal of Agricultural Research, 14(6):742–754.

Acknowledgements

The authors acknowledge financial support of Iowa State University Plant Sciences Institute Scholars Program, the Baker Center for Bioinformatics and Biological Statistics, and the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa, Project No. IOW03617, which is supported by USDA/NIFA and State of Iowa funds.
Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the U.S. Department of Agriculture.

Thank You!

Decomposition of $\mathbf{B}_j$ and $\mathbf{D}_j$

- Then the spectral decomposition of $\mathbf{W}_j$ is given by:

$$(\mathbf{N}_{r_j} \otimes \mathbf{N}_{c_j})\mathbf{W}_j(\mathbf{N}_{r_j}^{\mathrm{T}} \otimes \mathbf{N}_{c_j}^{\mathrm{T}}) = \theta_j \boldsymbol{\Lambda}_{r_j} \otimes \mathbf{I}_{c_j} + \bar{\theta}_j \mathbf{I}_{r_j} \otimes \boldsymbol{\Lambda}_{c_j}.$$

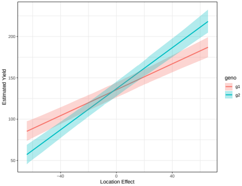
- $\boldsymbol{\Lambda}_k$ denote the $k \times k$ diagonal matrix whose uth diagonal entry is $4\sin^2\{\pi(u-1)/(2k)\}$.

- $\mathbf{N}_k$ denotes the $k \times k$ orthogonal matrix whose (u, v)th entry is $1/\sqrt{k}$ if $u = 1$, $\forall v$, and $(2/k)^{1/2}\cos\{\pi(u-1)(v-1/2)/k\}$ otherwise.

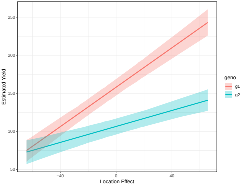
- $\mathbf{B}_j^{\mathrm{T}}$ denotes the $(r_j c_j - 1) \times r_j c_j$ matrix consisting of last $r_j c_j - 1$ rows of $\mathbf{N}_{r_j} \otimes \mathbf{N}_{c_j}$.

- $\mathbf{D}_j$ denotes the diagonal matrix consisting of the nonzero elements of $\theta_j \boldsymbol{\Lambda}_{r_j} \otimes \mathbf{I}_{c_j} + \bar{\theta}_j \mathbf{I}_{r_j} \otimes \boldsymbol{\Lambda}_{c_j}$.

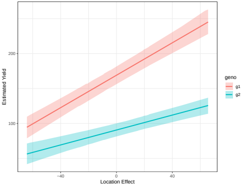
Assessing Uncertainty about FW Regression Lines



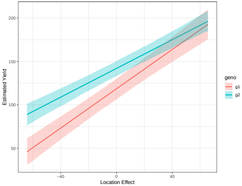
(a)



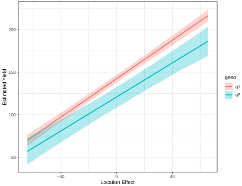
(b)



(c)



(d)



(e)

Figure: Estimated Yield vs Location Effect for pairs of genotypes

Model Evaluation via Within-Field Prediction

Level of spatial correlation vs performance of SFW model.