Spatial importance weighted cross validation

John Paige

NSM

June 19, 2024

Collaborators

Bob O'Hara

Focusing on a spatial setting

- Dataset: $\mathcal{X} = \{(s_1, y_1), \ldots, (s_n, y_n)\}$
- Spatial locations: $s_1, \ldots, s_n \overset{iid}{\sim} g(s)$ on spatial domain $\mathcal{D}$
- Response model:

$$y_i = h(s_i) + \epsilon_i, \;\; i = 1, \ldots, n.$$

- 'True signal' h is unknown but fixed, sufficiently well-behaved
- $\epsilon_1, \ldots, \epsilon_n$: independent, mean zero, $\text{Var}(\epsilon_i) = \sigma^2(s_i)$

How can we tell which models are best?

What do we mean by ‘best’?

How can we tell which models are best?

What do we mean by ‘best’?

- Inference vs prediction

How can we tell which models are best?

What do we mean by ‘best’?

- Inference vs **prediction**

How can we tell which models are best?

What do we mean by 'best'?

- Inference vs **prediction**
- Interpolation vs extrapolation

How can we tell which models are best?

What do we mean by 'best'?

- Inference vs **prediction**
- **Interpolation** vs extrapolation

How can we tell which models are best?

What do we mean by 'best'?

- Inference vs **prediction**
- **Interpolation** vs extrapolation
- ...

How can we tell which models are best?

What do we mean by 'best'?

- Inference vs **prediction**
- **Interpolation** vs extrapolation
- ...

Often LOOCV or K-fold CV are used

LOOCV in space: Basics

If $Y(s)$ is an observation depending on spatial location $s \subset \mathcal{D}$:

$$\begin{aligned} \text{LOOCV} &= \frac{1}{n}\sum_{i=1}^{n} \mathcal{S}(y_i; P_{\mathcal{X}_{-i}}) \\ &\approx E_{\hat{g},\hat{g}}[\mathcal{S}(Y(s); P_{\mathcal{X}})] \\ &\approx E_{g,g}[\mathcal{S}(Y(s); P_{\mathcal{X}})], \end{aligned}$$

for:

- $\mathcal{S}$: scoring rule
- $\mathcal{X}$: dataset
- $g(s)$: spatial density of data locations
- $E_{f,g}$: prediction locations from f, data locations from g

Criticisms of LOO and K-fold CV in spatial contexts

There are many claims (Adin et al., 2024; Roberts et al., 2017) that LOO and K-fold CV:

- 'are often overly optimistic'

Criticisms of LOO and K-fold CV in spatial contexts

There are many claims (Adin et al., 2024; Roberts et al., 2017) that LOO and K-fold CV:

- 'are often overly optimistic'
- 'don't account for spatial correlation'

Criticisms of LOO and K-fold CV in spatial contexts

There are many claims (Adin et al., 2024; Roberts et al., 2017) that LOO and K-fold CV:

- 'are often overly optimistic'
- 'don't account for spatial correlation'

Yet it is also acknowledged (Rabinowicz and Rosset, 2022) that, if the train and test data come from the same distribution, LOO and K-fold CV:

- are approximately unbiased for test error

Criticisms of LOO and K-fold CV in spatial contexts

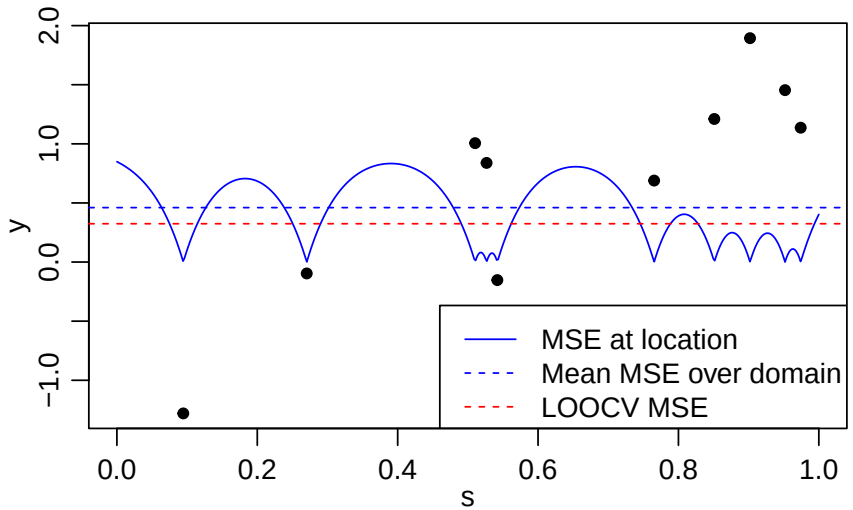
There are many claims (Adin et al., 2024; Roberts et al., 2017) that LOO and K-fold CV:

- 'are often overly optimistic'
- 'don't account for spatial correlation'

Yet it is also acknowledged (Rabinowicz and Rosset, 2022) that, if the train and test data come from the same distribution, LOO and K-fold CV:

- are approximately unbiased for test error
- are asymptotically consistent for test error

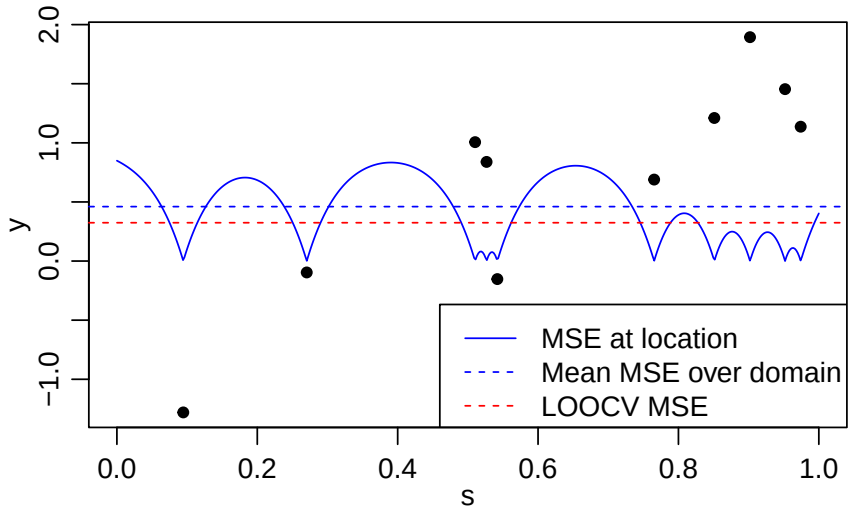
LOOCV in space: Thinking about nonrepresentativeness



If we oversample parts of the spatial domain, then:

- the data will tend to be in areas we oversample, and so

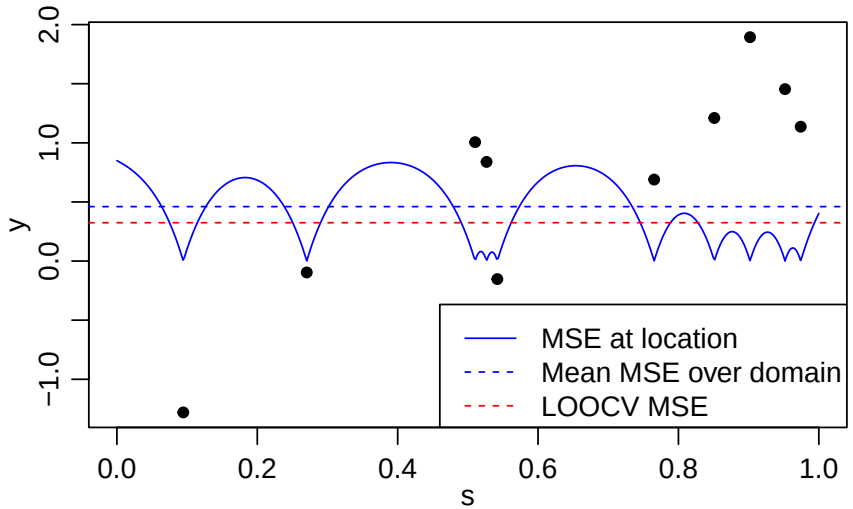
LOOCV in space: Thinking about nonrepresentativeness



If we oversample parts of the spatial domain, then:

- the data will tend to be in areas we oversample, and so
- the data will tend to be in areas where prediction errors are smaller than average, and so

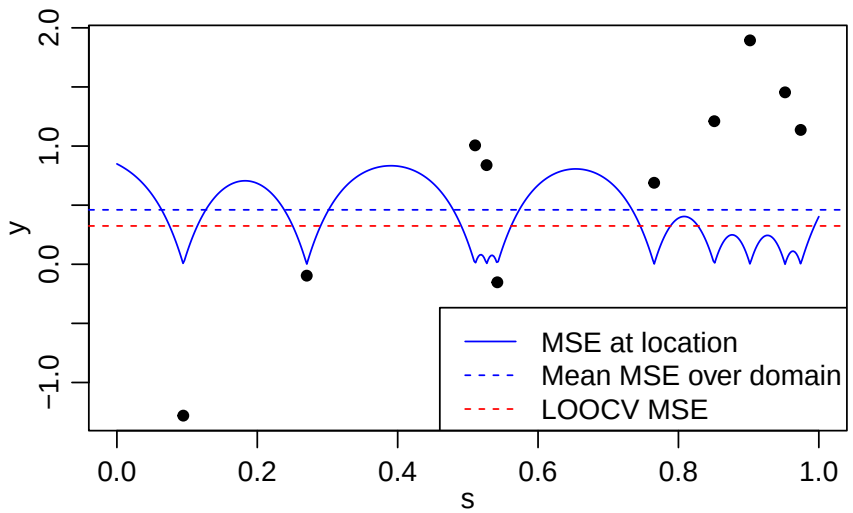
LOOCV in space: Thinking about nonrepresentativeness



If we oversample parts of the spatial domain, then:

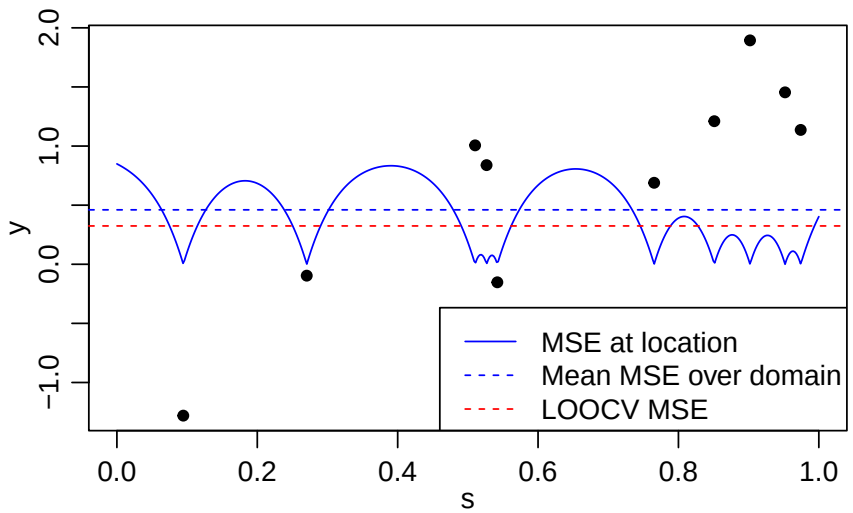
- the data will tend to be in areas we oversample, and so
- the data will tend to be in areas where prediction errors are smaller than average, and so
- LOOCV will be overly optimistic.

LOOCV in space: Thinking about nonrepresentativeness



- If we oversample parts of the spatial domain, then LOOCV is overly optimistic.
- Under representative sampling, LOOCV will be asymptotically unbiased regardless of spatial correlation (Rabinowicz and Rosset, 2022)

LOOCV in space: Thinking about nonrepresentativeness



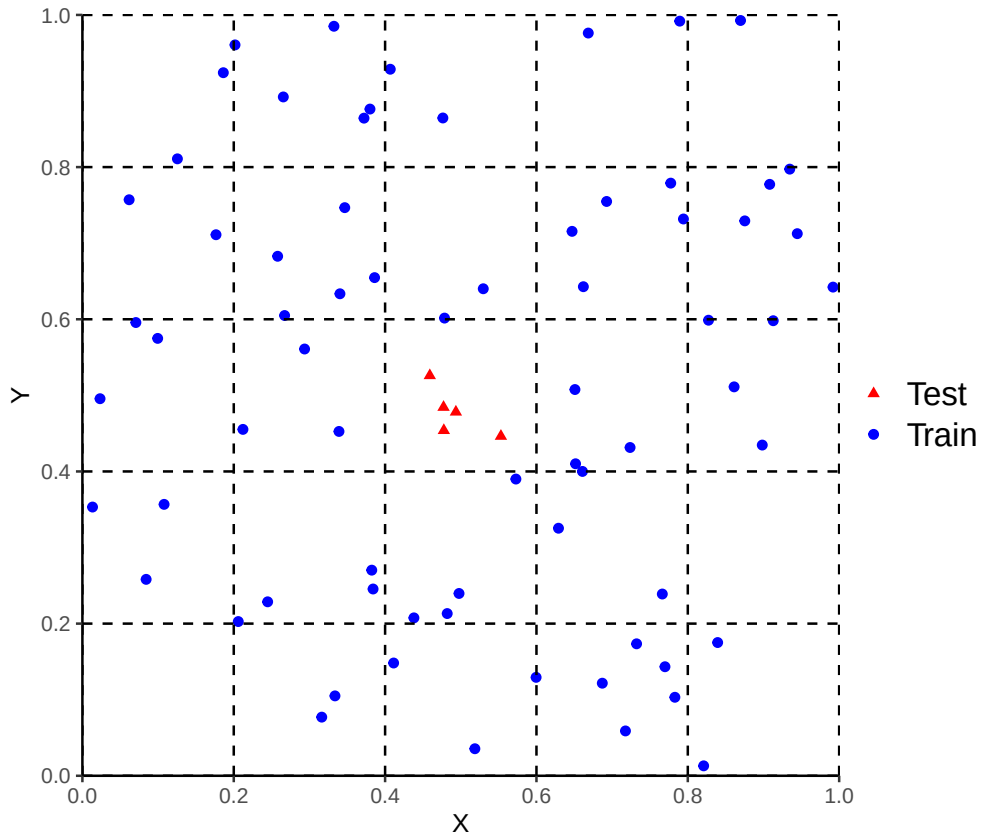
- If we oversample parts of the spatial domain, then LOOCV is overly optimistic.
- Under representative sampling, LOOCV will be asymptotically unbiased regardless of spatial correlation (Rabinowicz and Rosset, 2022)
- **How can we account for nonrepresentativeness in CV methods in space?**

Spatial CV methods

- Roberts et al. (2017) is a nice, if slightly outdated review of spatial CV methods:

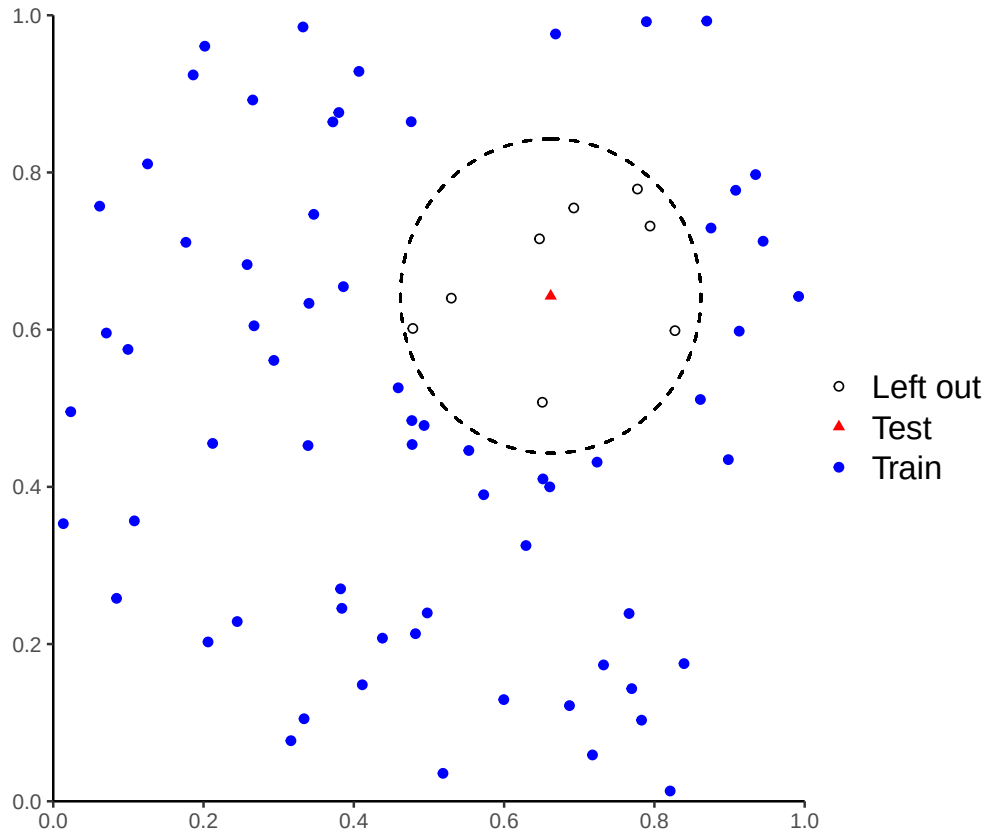
Spatial CV methods

- Roberts et al. (2017) is a nice, if slightly outdated review of spatial CV methods:
  - **Block CV**



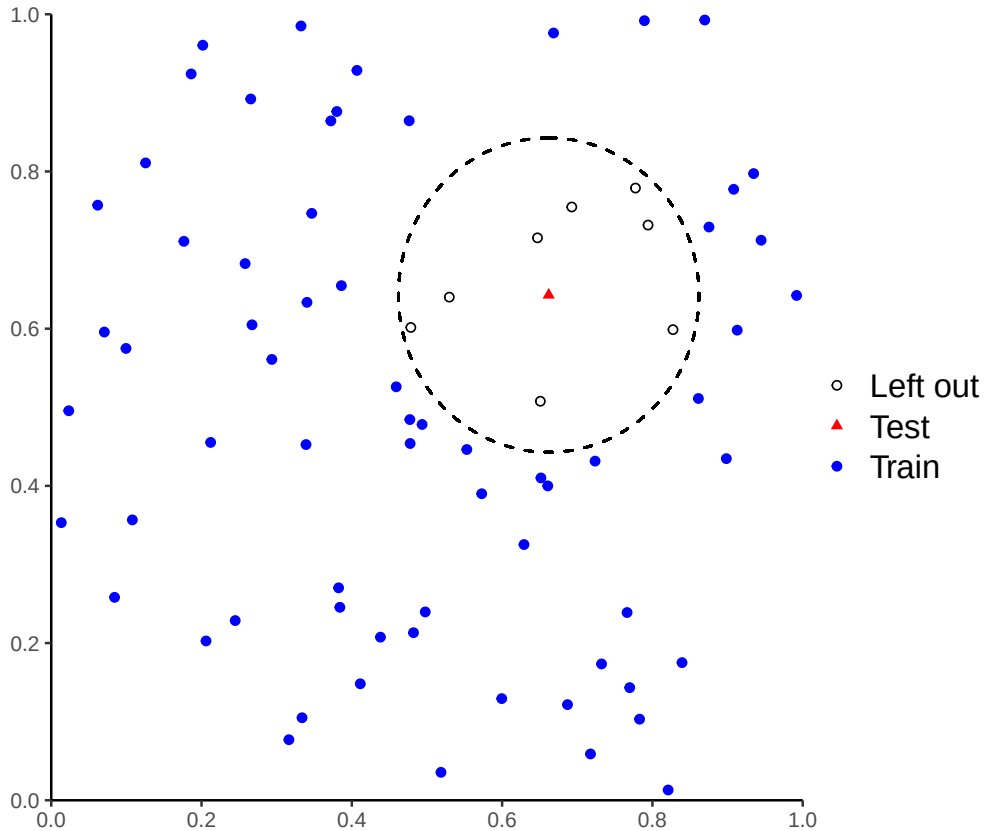
Spatial CV methods

- Roberts et al. (2017) is a nice, if slightly outdated review of spatial CV methods:
  - Block CV
  - **Buffered CV**



Spatial CV methods

- Roberts et al. (2017) is a nice, if slightly outdated review of spatial CV methods:
  - Block CV
  - Buffered CV
- **Leave group out CV (LGOCV)** Adin et al. (2024)



Defining interpolation error

Let $f(s) = \mathcal{U}(\mathcal{D})$ be the uniform distribution on $\mathcal{D}$. Then the expected interpolation error can be defined as:

$$E_{f,g}[\mathcal{S}(Y(s);\ P_{\mathcal{X}})] = \int_{\mathcal{D}} E[\mathcal{S}(Y(\tilde{s}); P_{\mathcal{X}})|s = \tilde{s}] \cdot f(\tilde{s})\ \mathrm{d}\tilde{s}$$
$$\neq E_{g,g}[\mathcal{S}(Y(s);\ P_{\mathcal{X}})] \qquad \text{(LOOCV)}$$

Rethinking block CV

Block CV with K grid cells, $G_1, \ldots, G_K$ of size $\delta > 0$ can be written:

$$\begin{aligned}
&\frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i \in \mathcal{I}^k} \mathcal{S}(y_i; P_{\mathcal{X}_{-\mathcal{I}^k}}) \\
&\approx \frac{1}{K} \sum_{k=1}^{K} \int_{G_k} E_g[\mathcal{S}(Y(\tilde{s}); P_{\mathcal{X}}) | s = \tilde{s}] \cdot \frac{1}{|G_k|} \, \mathrm{d}\tilde{s} \\
&= E_{f,g}[\mathcal{S}(Y(s); P_{\mathcal{X}})]
\end{aligned}$$

Rethinking block CV

Block CV with K grid cells, $G_1, \ldots, G_K$ of size $\delta > 0$ can be written:

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i\in\mathcal{I}^k}\mathcal{S}(y_i; P_{\mathcal{X}_{-\mathcal{I}^k}}) \\
&\approx \frac{1}{K}\sum_{k=1}^{K}\int_{G_k} E_g[\mathcal{S}(Y(\tilde{s}); P_{\mathcal{X}})|s=\tilde{s}]\cdot\frac{1}{|G_k|}\,\mathrm{d}\tilde{s} \\
&= E_{f,g}[\mathcal{S}(Y(s); P_{\mathcal{X}})]
\end{aligned}$$

Problems:

Rethinking block CV

Block CV with K grid cells, $G_1, \ldots, G_K$ of size $\delta > 0$ can be written:

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i\in\mathcal{I}^k}\mathcal{S}(y_i; P_{\mathcal{X}_{-\mathcal{I}^k}}) \\
&\approx \frac{1}{K}\sum_{k=1}^{K}\int_{G_k} E_g[\mathcal{S}(Y(\tilde{s}); P_{\mathcal{X}})|s=\tilde{s}]\cdot\frac{1}{|G_k|}\,\mathrm{d}\tilde{s} \\
&= E_{f,g}[\mathcal{S}(Y(s); P_{\mathcal{X}})]
\end{aligned}$$

Problems:

- What if some grid cells have no samples?

Rethinking block CV

Block CV with K grid cells, $G_1, \ldots, G_K$ of size $\delta > 0$ can be written:

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i\in\mathcal{I}^k}\mathcal{S}(y_i; P_{\mathcal{X}_{-\mathcal{I}^k}}) \\
&\approx \frac{1}{K}\sum_{k=1}^{K}\int_{G_k} E_g[\mathcal{S}(Y(\tilde{s}); P_{\mathcal{X}})|s=\tilde{s}]\cdot\frac{1}{|G_k|}\,\mathrm{d}\tilde{s} \\
&= E_{f,g}[\mathcal{S}(Y(s); P_{\mathcal{X}})]
\end{aligned}$$

Problems:

- What if some grid cells have no samples?
- How do you choose the grid resolution? Too small $\rightarrow$ LOOCV, too large $\rightarrow$ too much extrapolation

Re-Rethinking block CV: Voronoi Cell CV

Let $V_1, \ldots, V_n \subset \mathcal{D}$ be Voronoi cells generated by observation locations $s_1, \ldots, s_n$. Then:

$$E_{f,g}[\mathcal{S}(Y(s); P_{\mathcal{X}})]$$

Re-Rethinking block CV: Voronoi Cell CV

Let $V_1, \ldots, V_n \subset \mathcal{D}$ be Voronoi cells generated by observation locations $s_1, \ldots, s_n$. Then:

$$
\begin{aligned}
&E_{f,g}[\mathcal{S}(Y(s); P_{\mathcal{X}})] \\
&= \sum_{i=1}^{n} \int_{V_i} E[\mathcal{S}(Y(\tilde{s}); P_{\mathcal{X}}) | s = \tilde{s}] \cdot \frac{1}{|\mathcal{D}|} \, \mathrm{d}\tilde{s}
\end{aligned}
$$

Re-Rethinking block CV: Voronoi Cell CV

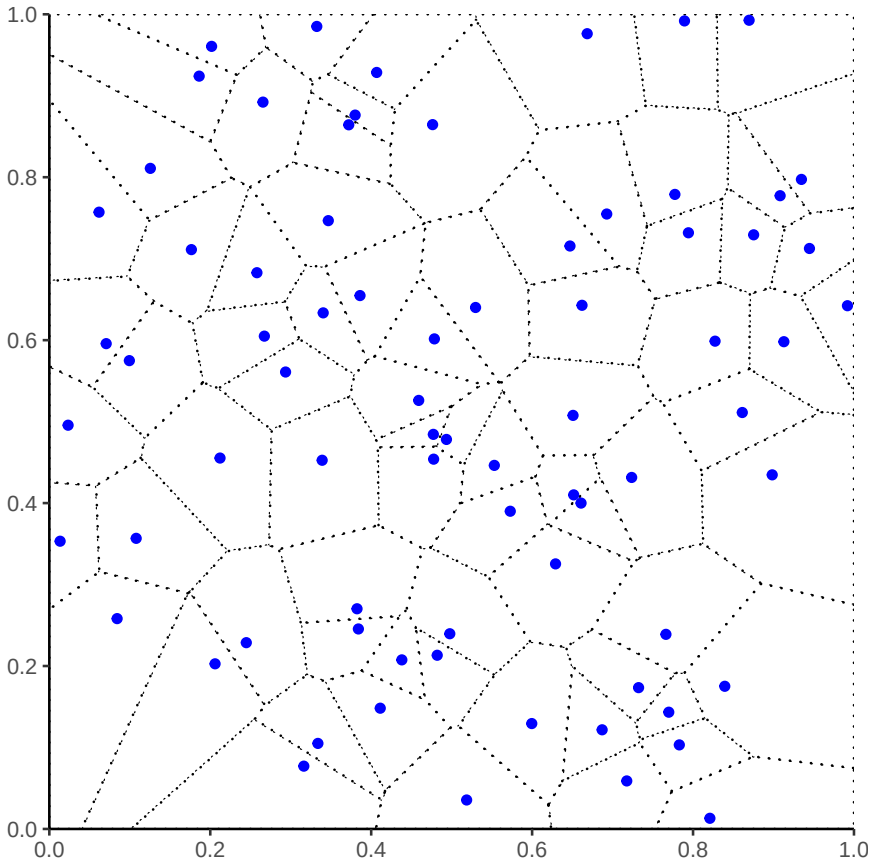
Let $V_1, \ldots, V_n \subset \mathcal{D}$ be Voronoi cells generated by observation locations $s_1, \ldots, s_n$. Then:

$$
\begin{aligned}
& E_{f,g}[\mathcal{S}(Y(s); P_{\mathcal{X}})] \\
&= \sum_{i=1}^{n} \int_{V_i} E[\mathcal{S}(Y(\tilde{s}); P_{\mathcal{X}}) | s = \tilde{s}] \cdot \frac{1}{|\mathcal{D}|} \, \mathrm{d}\tilde{s} \\
&\approx \sum_{i=1}^{n} \frac{|V_i|}{|\mathcal{D}|} \mathcal{S}(y_i; P_{\mathcal{X}_{-i}}).
\end{aligned}
$$

- $|V_i|/|\mathcal{D}|$ is asymptotically unbiased for $f(s_i)/g(s_i)$, an importance weight (Barr and Schoenberg, 2010)
- This is importance weighted CV (IWCV, Sugiyama et al. 2007), domain adaptation (Farahani et al., 2021)

Voronoi Cell CV

$$\text{VCCV} = \sum_{i=1}^{n} \frac{|V_i|}{|\mathcal{D}|} \mathcal{S}(y_i; P_{\mathcal{X}_{-i}}).$$



Nonstationary error variance: Problem setup

- A random 30% of $\mathcal{D}$ is mountains and 70% is plains

Nonstationary error variance: Problem setup

- A random 30% of $\mathcal{D}$ is mountains and 70% is plains
- **True model:** $h(s)$ is a realization of a mean 0 Gaussian random field with Matérn covariance with parameters $\rho = 0.2$, $\sigma^2 = 1$, $\nu = 2.5$.

$$\text{Var}(\sigma^2_\epsilon(s)) = \begin{cases} 1 & s \text{ in mountains,} \\ 0.1^2 & s \text{ in plains.} \end{cases}$$

Nonstationary error variance: Problem setup

- A random 30% of $\mathcal{D}$ is mountains and 70% is plains
- **True model:** $h(s)$ is a realization of a mean 0 Gaussian random field with Matérn covariance with parameters $\rho = 0.2$, $\sigma^2 = 1$, $\nu = 2.5$.

$$\text{Var}(\sigma^2_\epsilon(s)) = \begin{cases} 1 & s \text{ in mountains,} \\ 0.1^2 & s \text{ in plains.} \end{cases}$$

- **Wrong model 1:** Best stationary model by LOOCV

Nonstationary error variance: Problem setup

- A random 30% of $\mathcal{D}$ is mountains and 70% is plains
- **True model:** $h(s)$ is a realization of a mean 0 Gaussian random field with Matérn covariance with parameters $\rho = 0.2$, $\sigma^2 = 1$, $\nu = 2.5$.

$$\text{Var}(\sigma^2_\epsilon(s)) = \begin{cases} 1 & s \text{ in mountains,} \\ 0.1^2 & s \text{ in plains.} \end{cases}$$

- **Wrong model 1:** Best stationary model by LOOCV
- **Wrong model 2:** Assumes error variance in mountains is σ^2_M and in plains is σ^2_P for some fixed values.

Nonstationary error variance: Problem setup

- A random 30% of $\mathcal{D}$ is mountains and 70% is plains
- **True model:** $h(s)$ is a realization of a mean 0 Gaussian random field with Matérn covariance with parameters $\rho = 0.2$, $\sigma^2 = 1$, $\nu = 2.5$.

$$\text{Var}(\sigma^2_\epsilon(s)) = \begin{cases} 1 & s \text{ in mountains,} \\ 0.1^2 & s \text{ in plains.} \end{cases}$$

- **Wrong model 1:** Best stationary model by LOOCV
- **Wrong model 2:** Assumes error variance in mountains is σ^2_M and in plains is σ^2_P for some fixed values.
- R: level of oversampling in the mountains

Interval score

- The interval score (Gneiting and Raftery, 2007) at significance level $\alpha \in (0, 1)$ is:

$$\begin{aligned} \mathsf{IS}_\alpha(Y_i; P_i) &= (u_i - l_i) + \frac{2}{\alpha}(l_i - Y_i) I\{Y_i < l_i\} \\ &\quad + \frac{2}{\alpha}(Y_i - u_i) I\{Y_i > u_i\} \\ \mathsf{LOO} &= \frac{1}{n} \sum_{i=1}^{n} \mathsf{IS}_\alpha(Y_i; P_i) \\ \mathsf{IW} &= \sum_{i=1}^{n} w_i \mathsf{IS}_\alpha(Y_i; P_i) \end{aligned}$$

Scoring models

Let

$$\Delta^{\text{LOO}} = \text{LOO}_1 - \text{LOO}_2, \qquad \Delta^{\text{IW}} = \text{IW}_1 - \text{IW}_2, \qquad \text{and}$$

$$\Delta^{\mathcal{D}} = E_{g,g}[\Delta^{\text{IW}}] = E_{f,f}[\Delta^{\text{LOO}}].$$

Under infill asymptotics, use the scores/metrics:

- Probability of selecting the best model

Scoring models

Let

$$\Delta^{\text{LOO}} = \text{LOO}_1 - \text{LOO}_2, \qquad \Delta^{\text{IW}} = \text{IW}_1 - \text{IW}_2, \qquad \text{and}$$

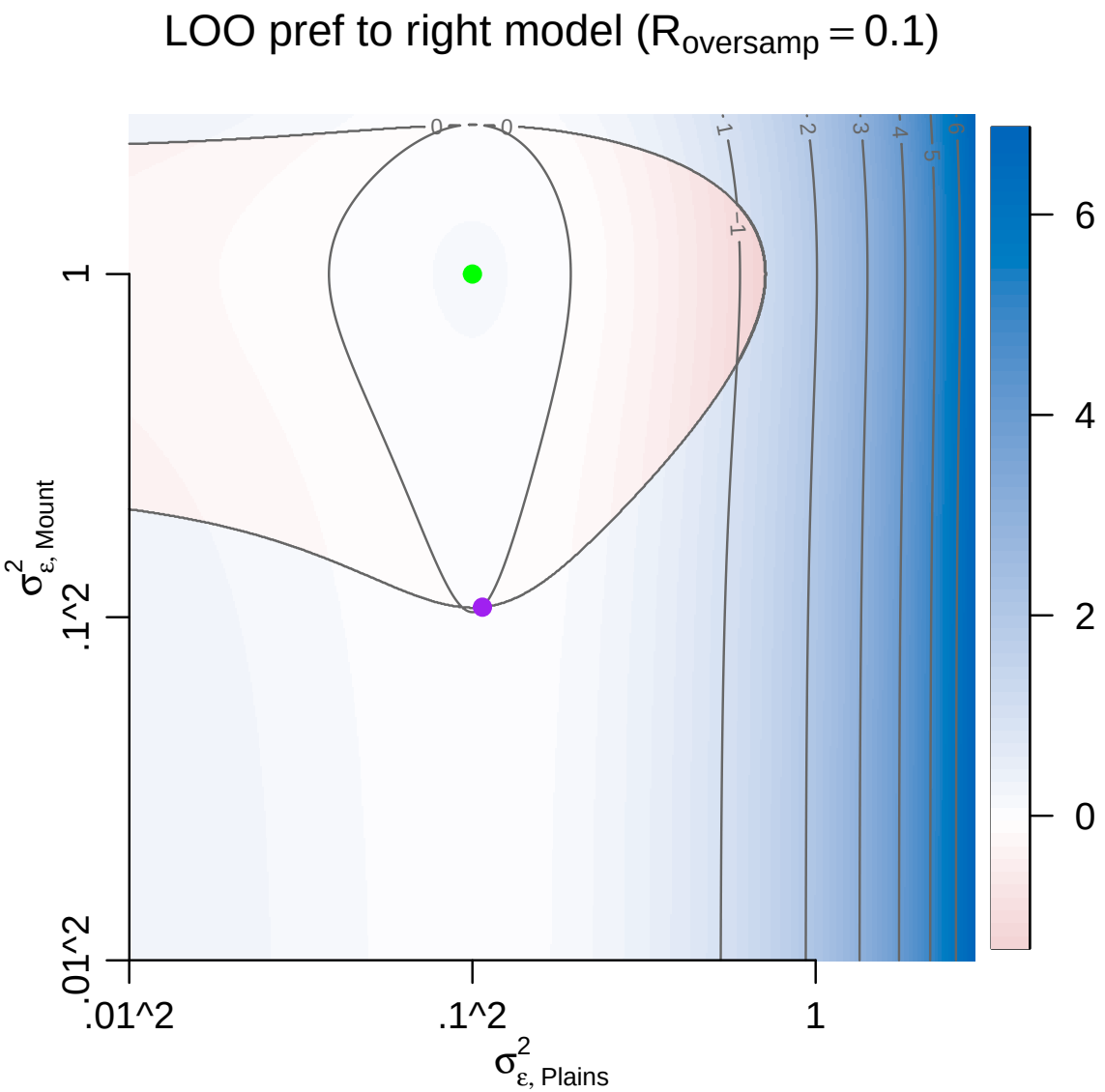
$$\Delta^{\mathcal{D}} = E_{g,g}[\Delta^{\text{IW}}] = E_{f,f}[\Delta^{\text{LOO}}].$$

Under infill asymptotics, use the scores/metrics:

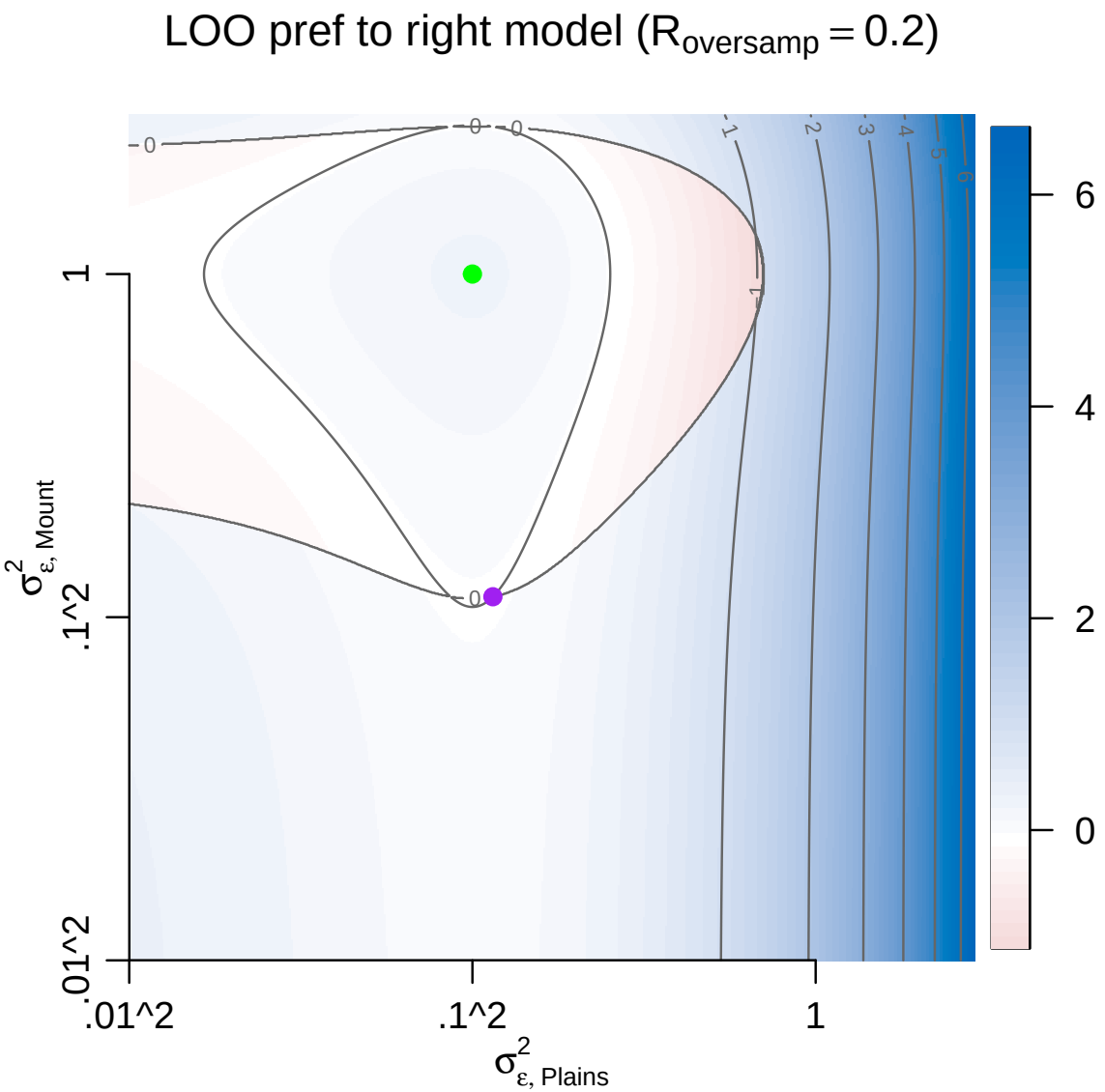
- Probability of selecting the best model
- Mean preference towards best model:

$$|E[\Delta^{\text{LOO}}]|\text{sign}\{E[\Delta^{\text{LOO}}] \cdot E[\Delta^{\mathcal{D}}]\}$$

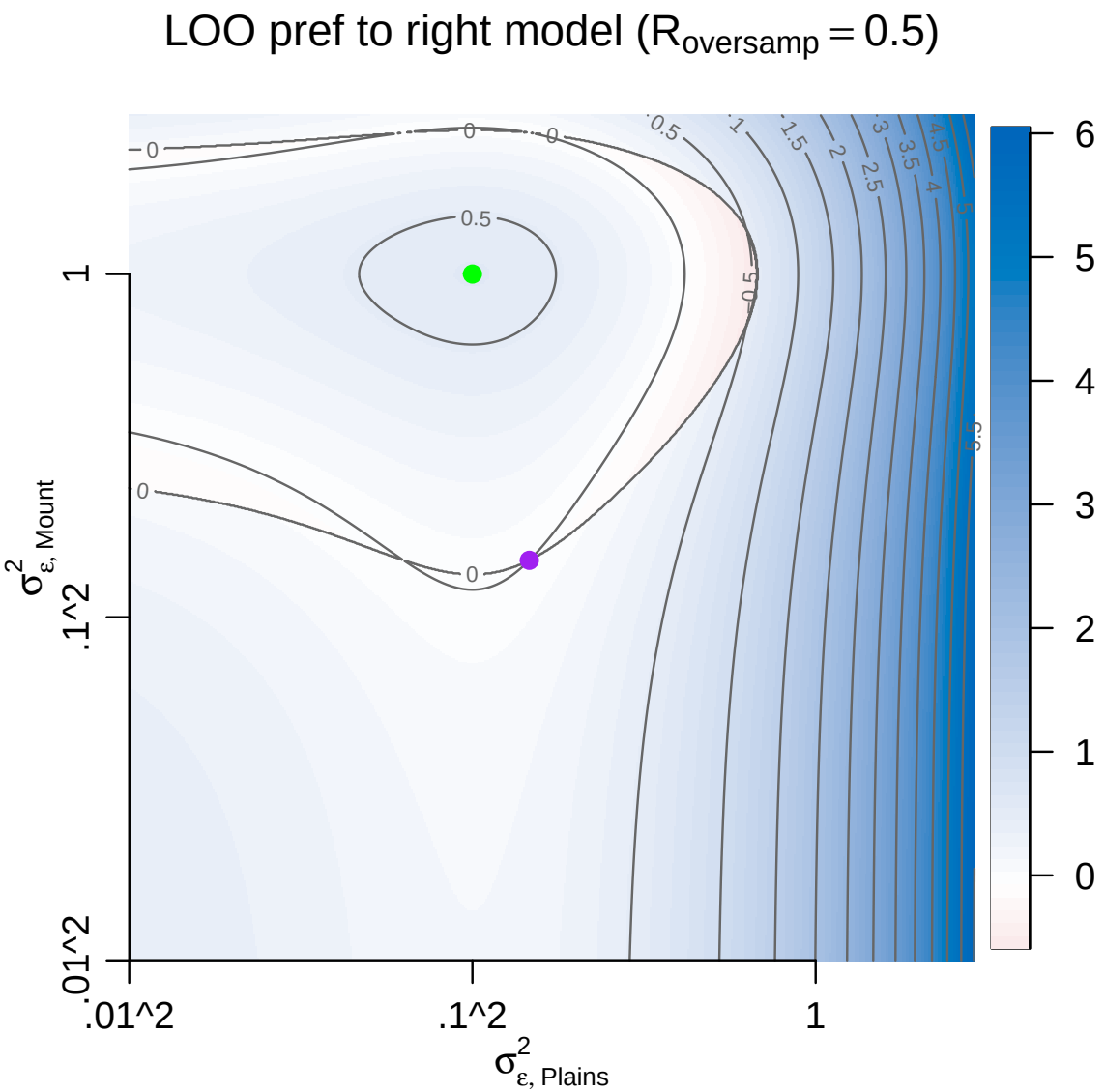
LOO preference of best model vs nonrepresentativeness



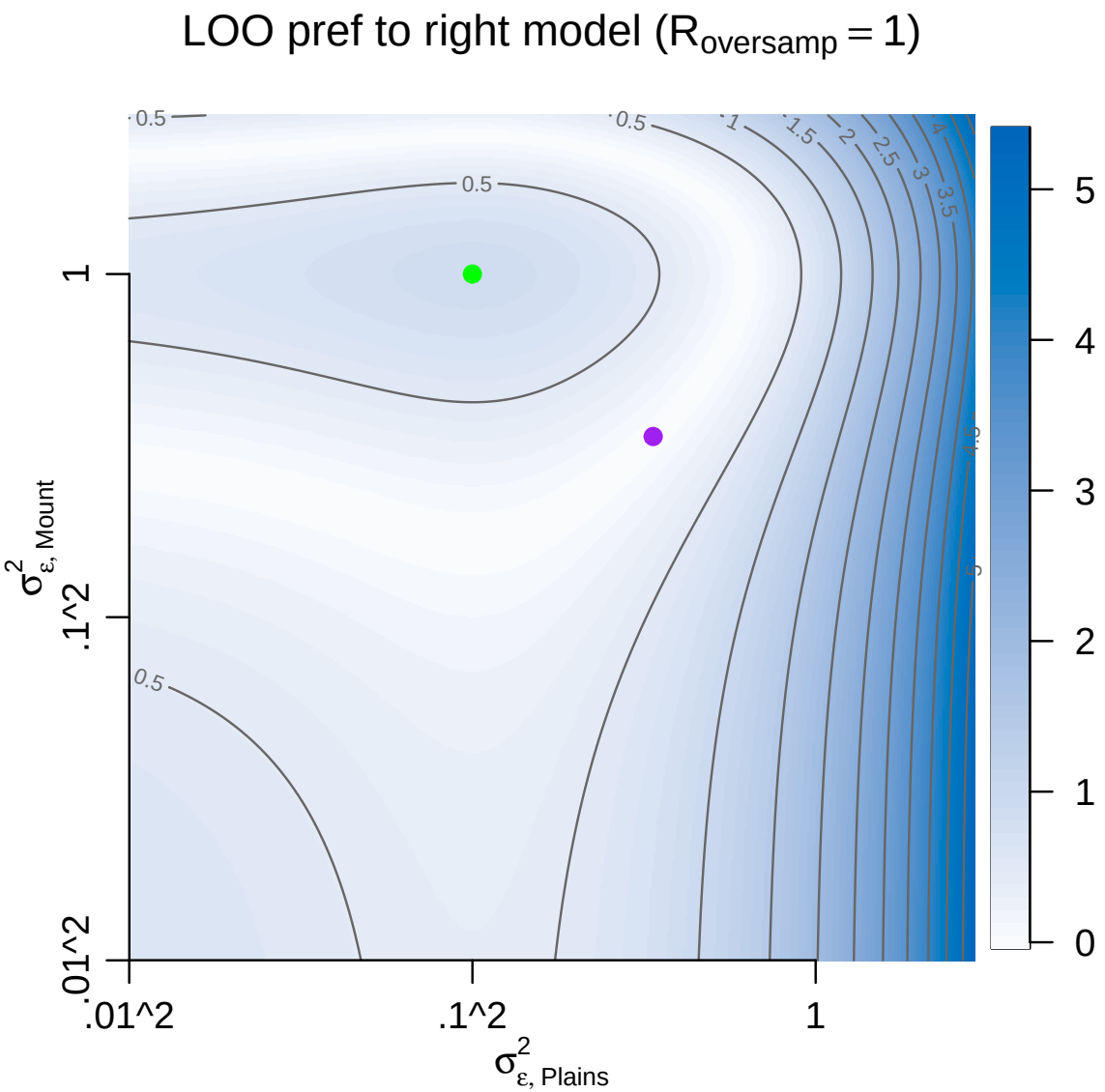
LOO preference of best model vs nonrepresentativeness



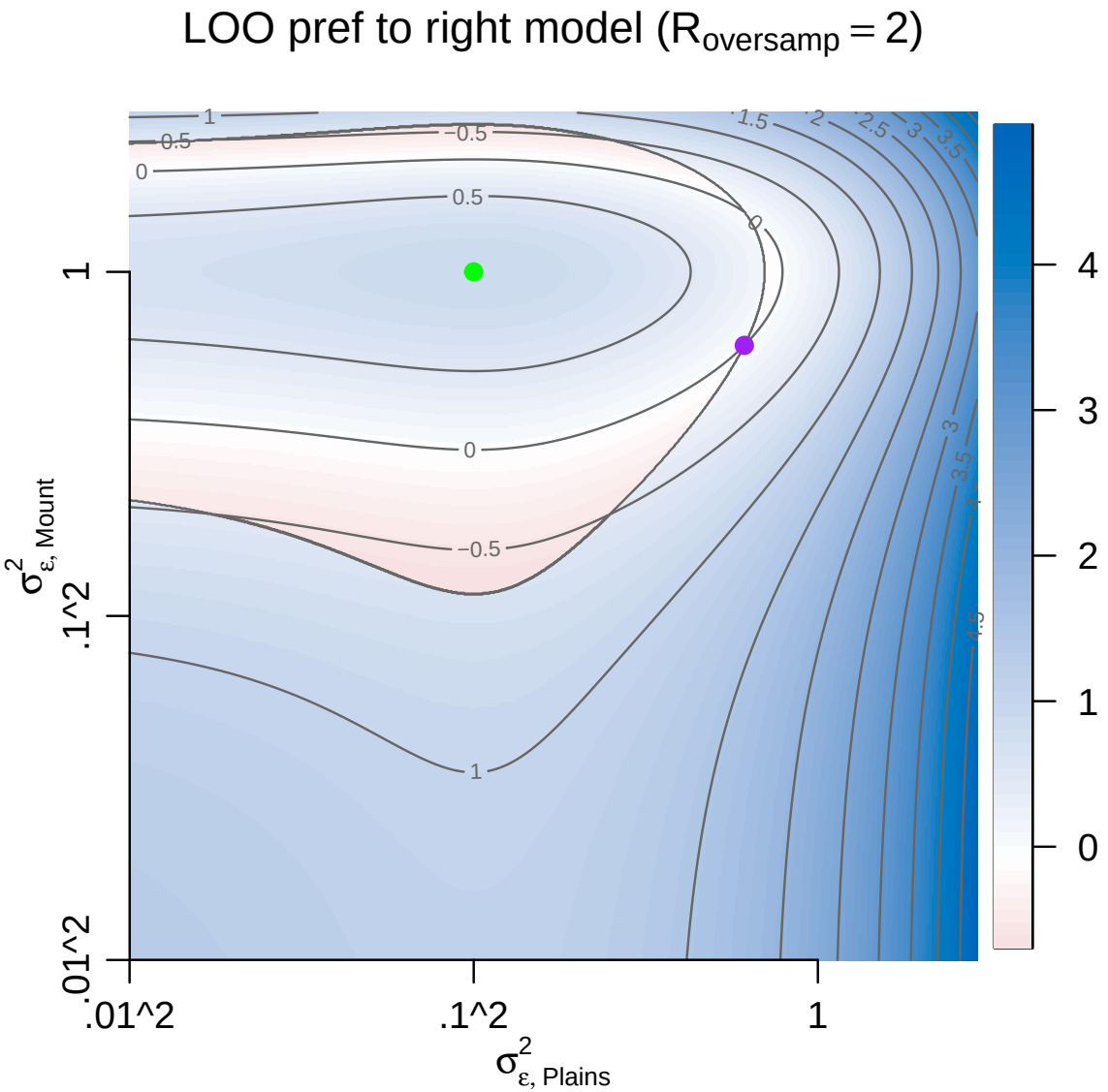
LOO preference of best model vs nonrepresentativeness



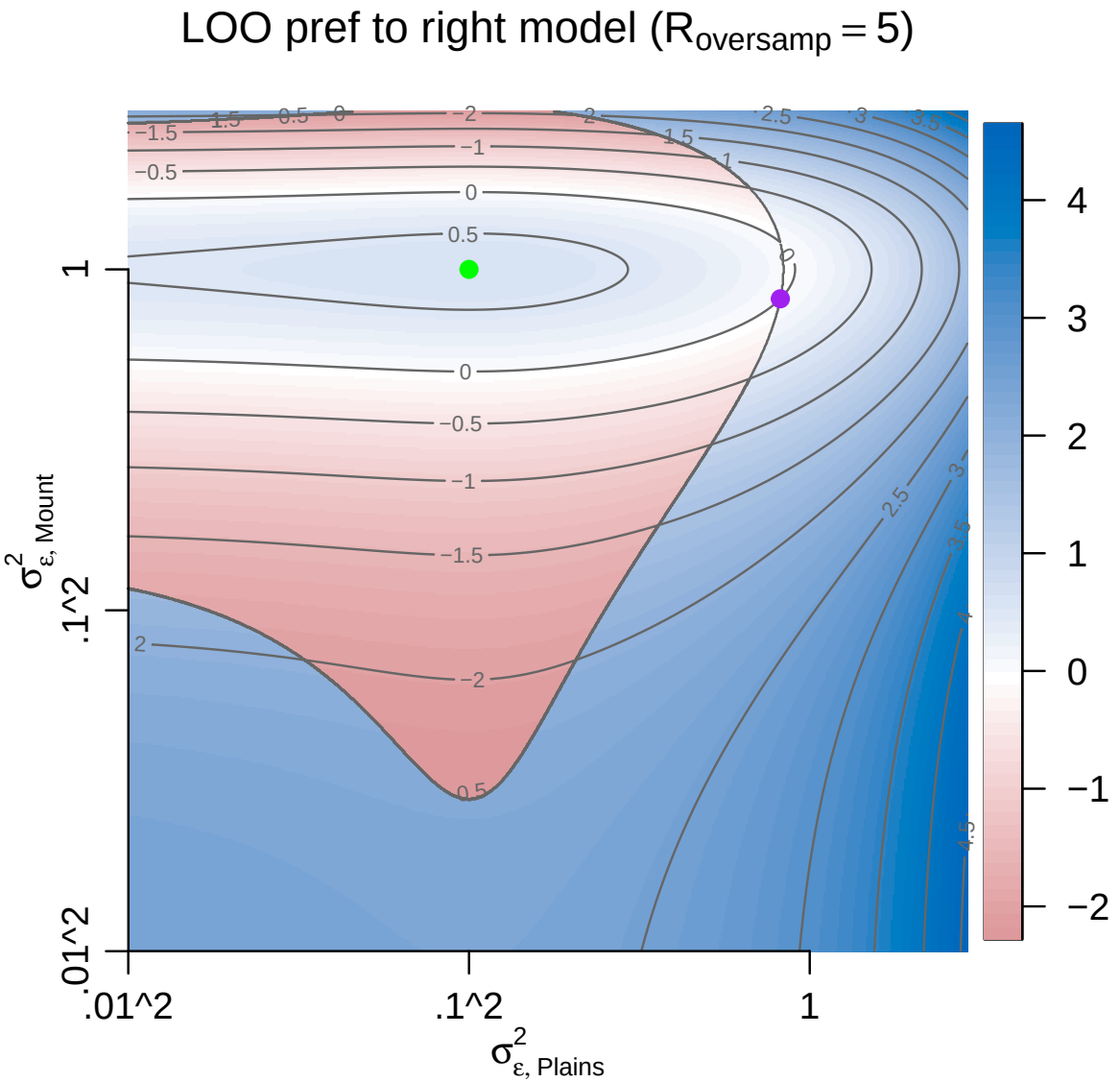
LOO preference of best model vs nonrepresentativeness



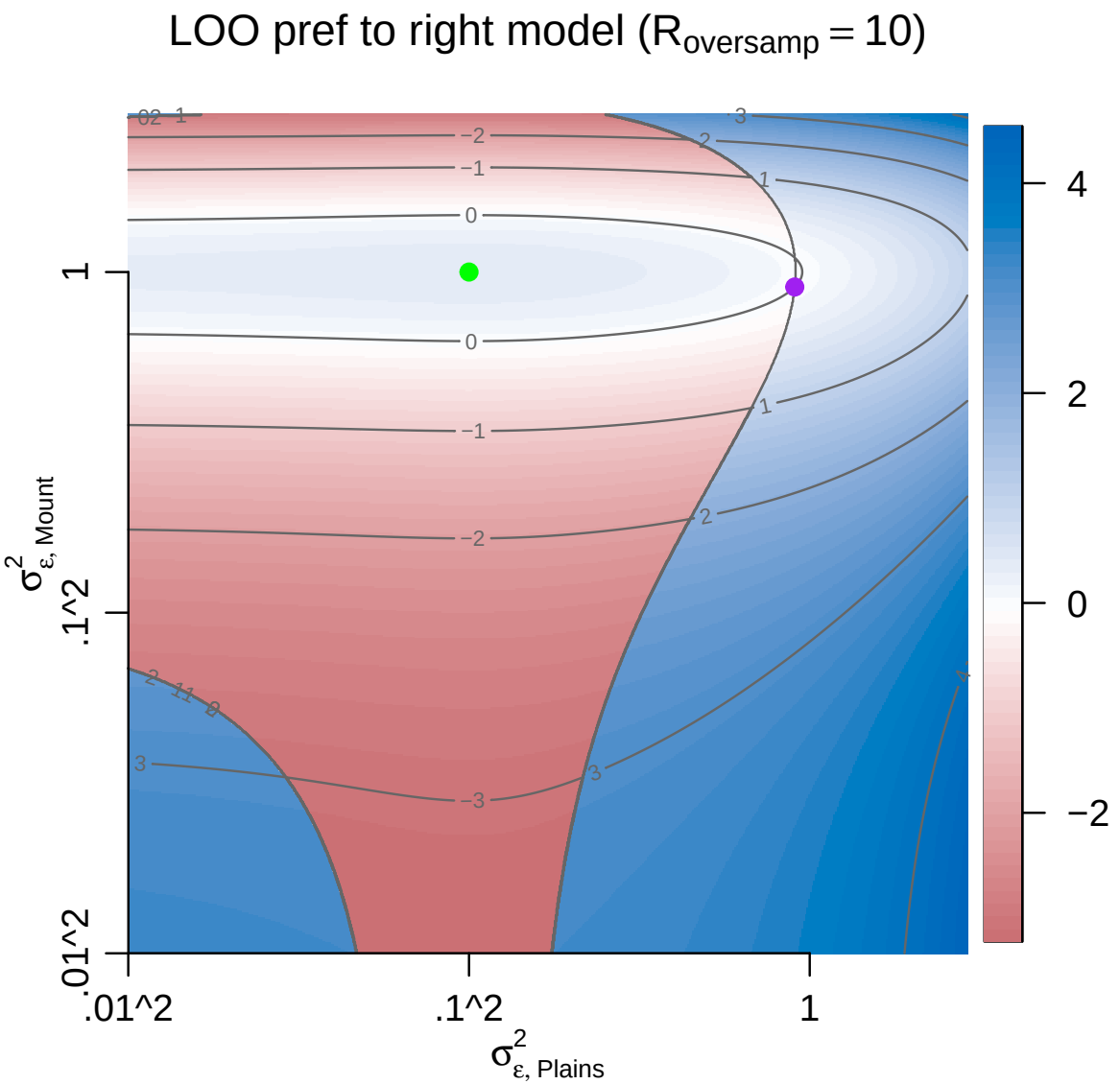
LOO preference of best model vs nonrepresentativeness



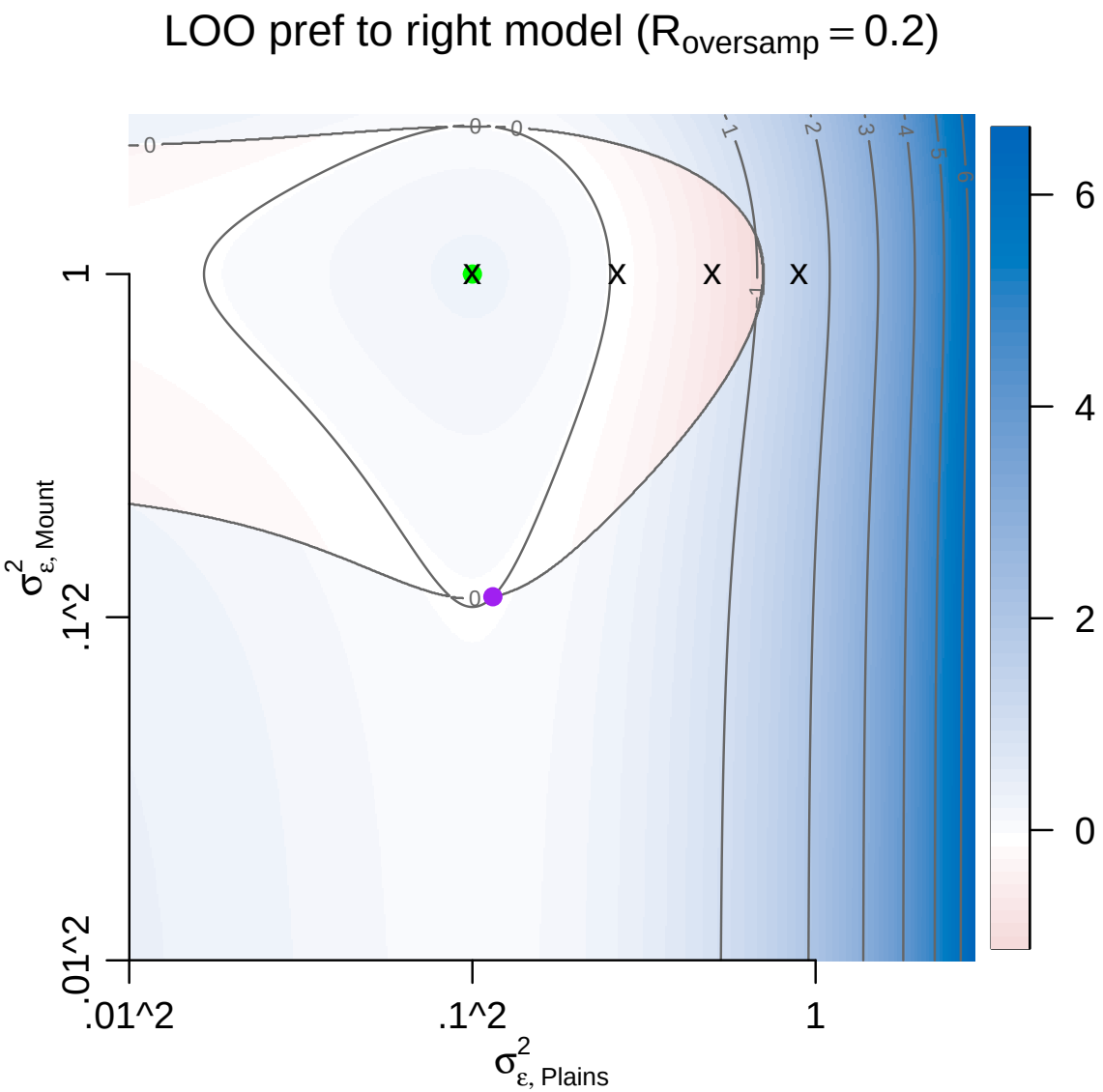
LOO preference of best model vs nonrepresentativeness



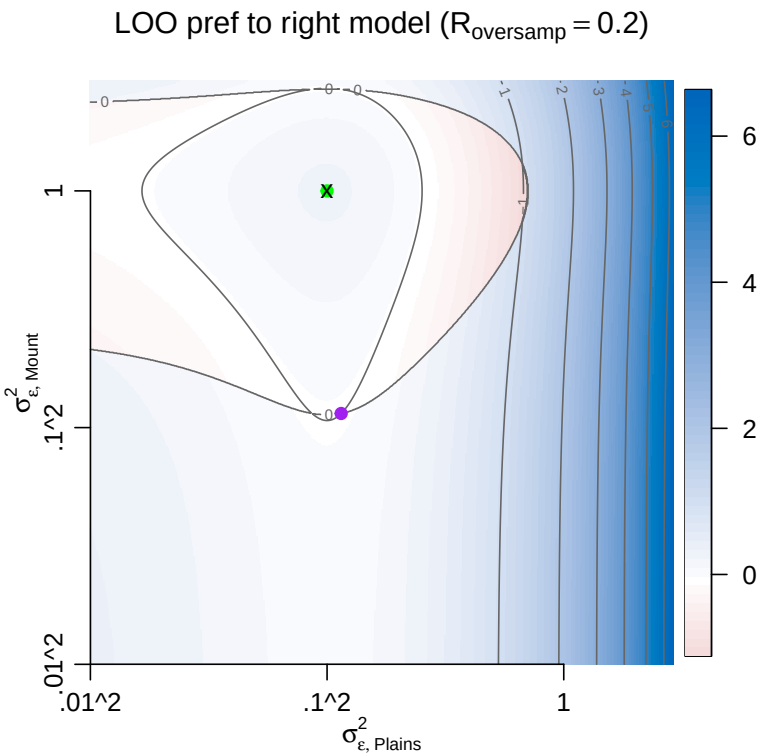
LOO preference of best model vs nonrepresentativeness

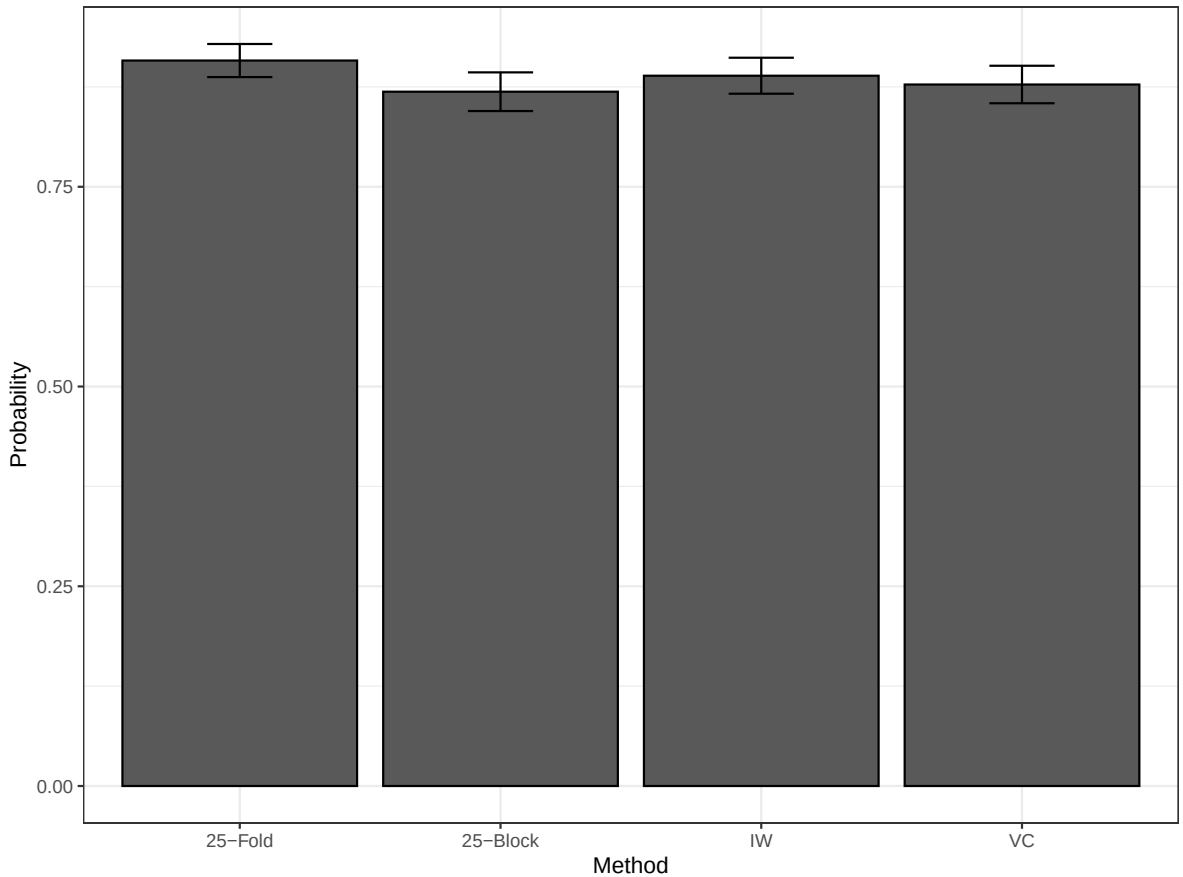


Nonstationary error variance: Simulations

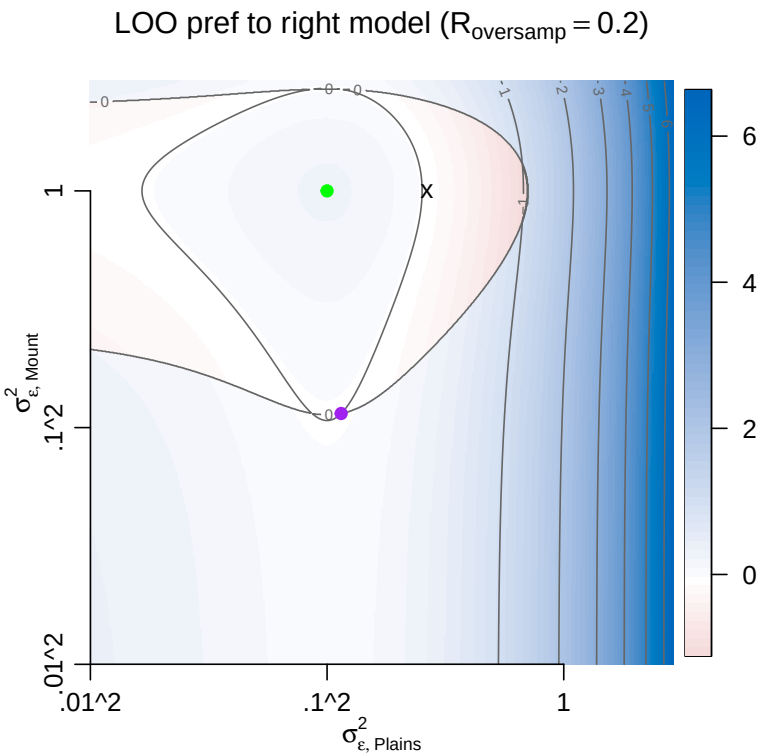


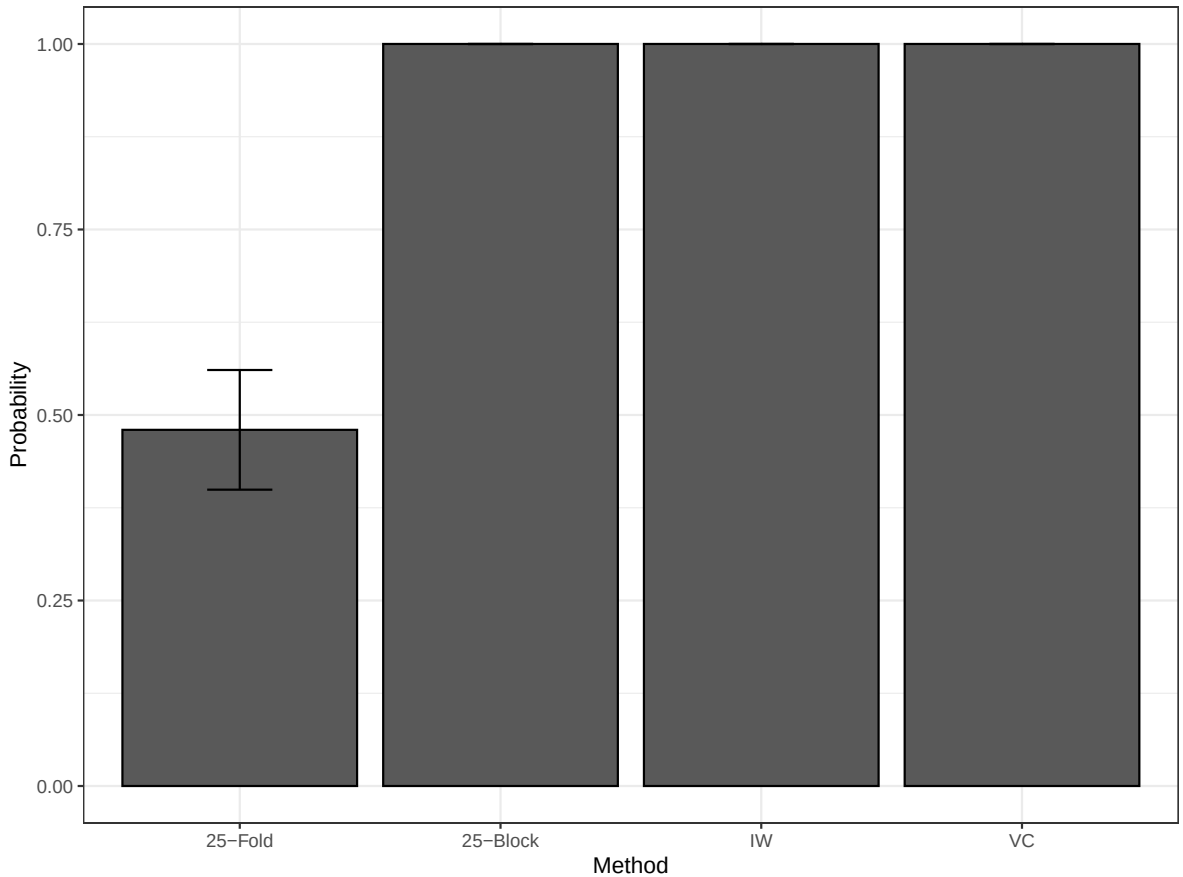
Scenario 1: Comparison vs true model
$n = 50$, 25-fold CV





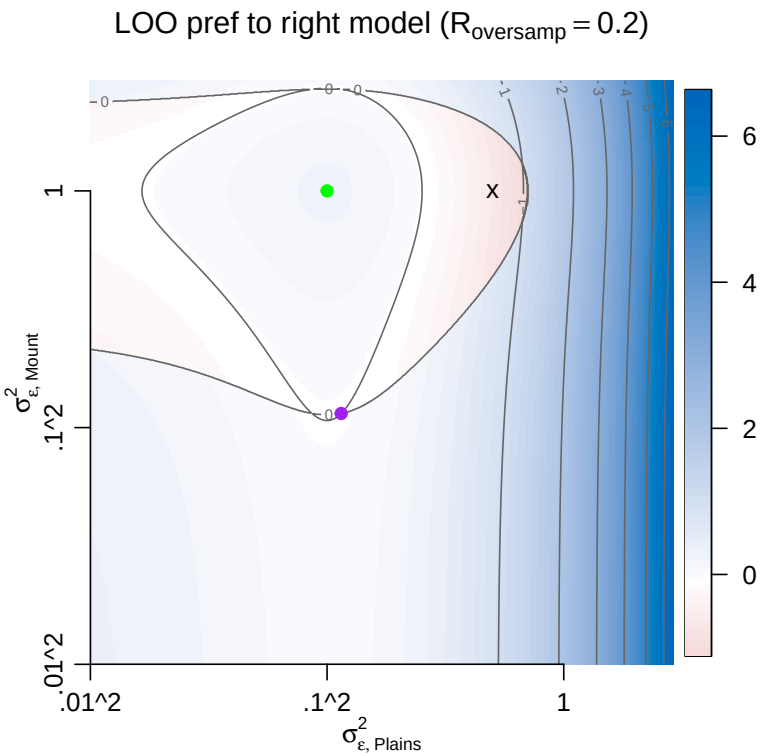
Scenario 2: Equivalent LOO models
$n = 2000$, 25-fold CV

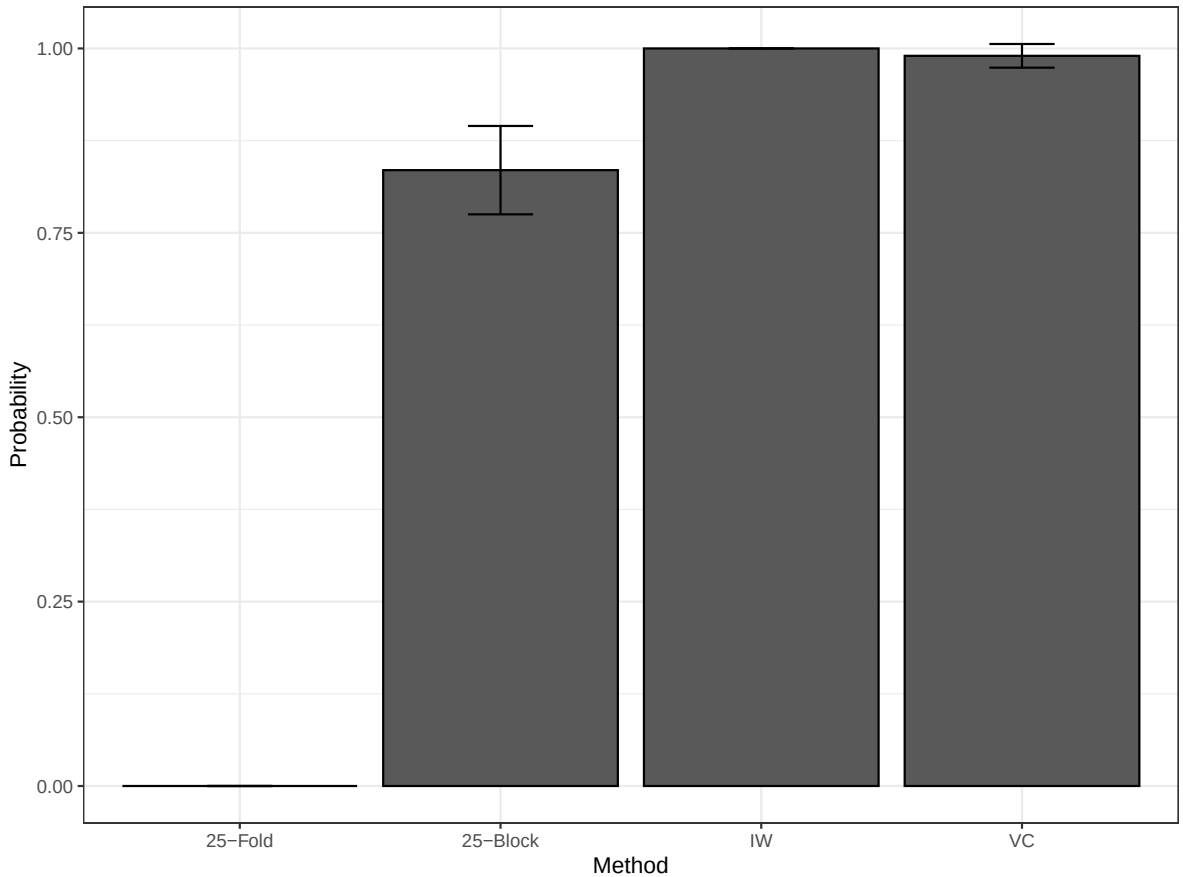




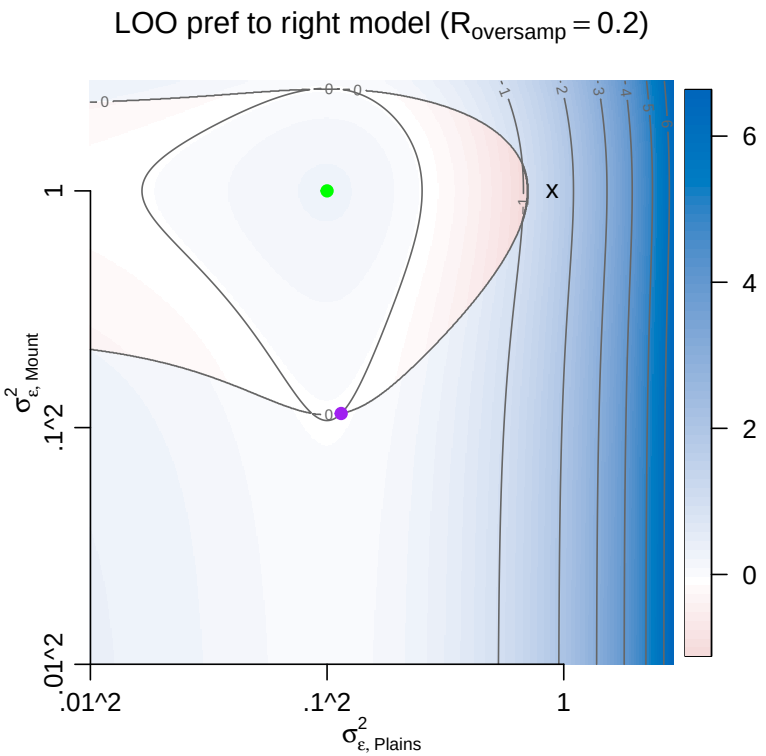
Scenario 3: Better model, worse under LOO

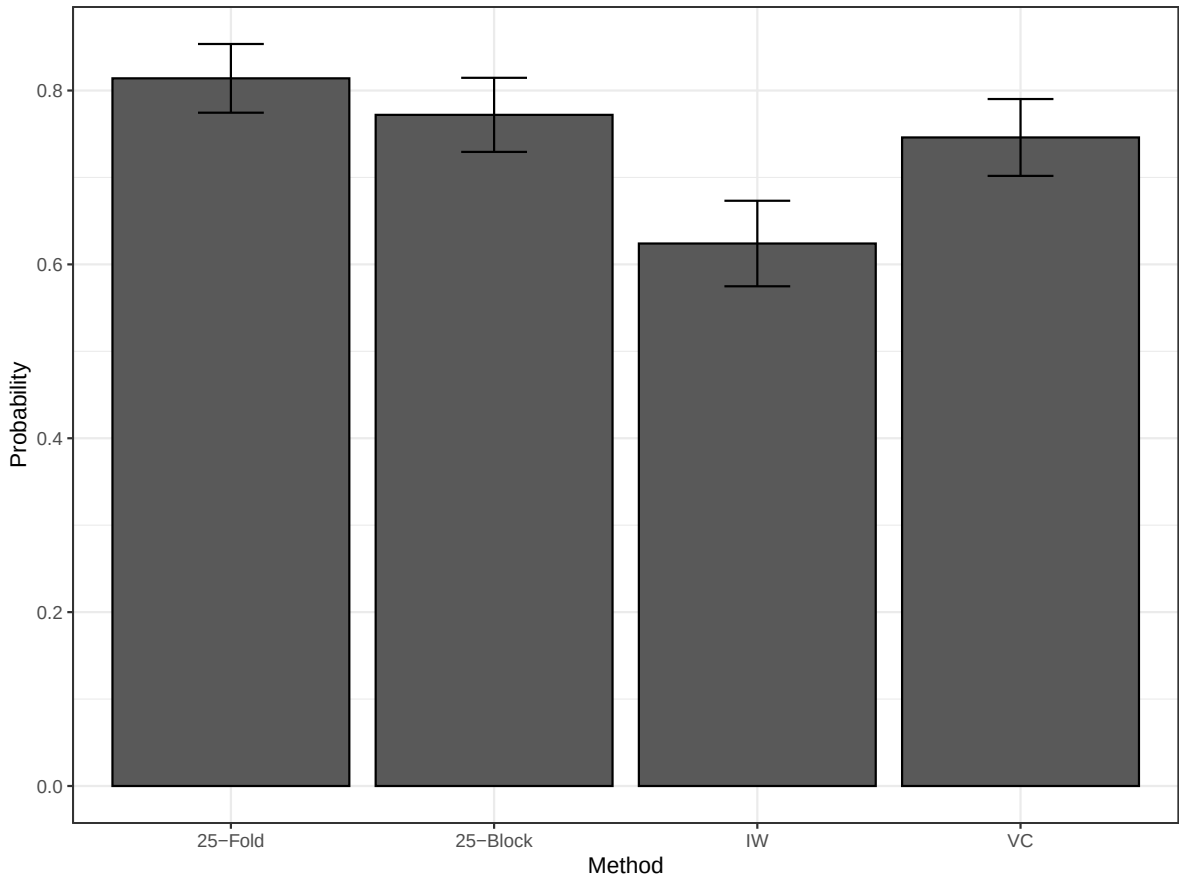
$n = 2000$, 25-fold CV





Scenario 4: Advantage of LOO's lower variance
$n = 200$, 25-fold CV





Conclusions

- LOO and K-fold CV can work well in geostatistical settings under representative spatial sampling

Conclusions

- LOO and K-fold CV can work well in geostatistical settings under representative spatial sampling
- Nonrepresentative sampling and nonstationary error variance cause problems for LOO and K-fold CV:
  - High bias and MSE
  - Can select wrong model with probability $\to 1$ as $n \to \infty$

Conclusions

- LOO and K-fold CV can work well in geostatistical settings under representative spatial sampling
- Nonrepresentative sampling and nonstationary error variance cause problems for LOO and K-fold CV:
  - High bias and MSE
  - Can select wrong model with probability $\to 1$ as $n \to \infty$
- VC CV is more robust, and is consistent for the best model in all considered scenarios

Conclusions

- LOO and K-fold CV can work well in geostatistical settings under representative spatial sampling
- Nonrepresentative sampling and nonstationary error variance cause problems for LOO and K-fold CV:
  - High bias and MSE
  - Can select wrong model with probability $\to 1$ as $n \to \infty$
- VC CV is more robust, and is consistent for the best model in all considered scenarios
- VC CV at times performs better than IW CV due to its accounting for sampling variability

Conclusions

- LOO and K-fold CV can work well in geostatistical settings under representative spatial sampling
- Nonrepresentative sampling and nonstationary error variance cause problems for LOO and K-fold CV:
  - High bias and MSE
  - Can select wrong model with probability $\to 1$ as $n \to \infty$
- VC CV is more robust, and is consistent for the best model in all considered scenarios
- VC CV at times performs better than IW CV due to its accounting for sampling variability
- VC CV is simple and requires no tuning parameters

Limitations, potential future work

- No ‘one size fits all’ solution
- Extension to space-time
- Methods for variance reduction, weight smoothing

References I

Adin, A., E. T. Krainski, A. Lenzi, Z. Liu, J. Martínez-Minaya, and H. Rue (2024). Automatic cross-validation in structured models: Is it time to leave out leave-one-out? *Spatial Statistics*, 100843.

Barr, C. D. and F. P. Schoenberg (2010). On the Voronoi estimator for the intensity of an inhomogeneous planar Poisson process. *Biometrika 97*, 977–984.

Farahani, A., S. Voghoei, K. Rasheed, and H. R. Arabnia (2021). A brief review of domain adaptation. In *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894. Springer.

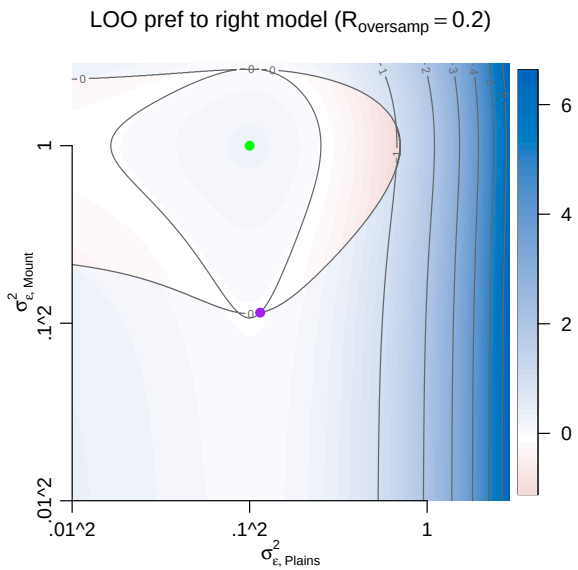
Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*, 359–378.

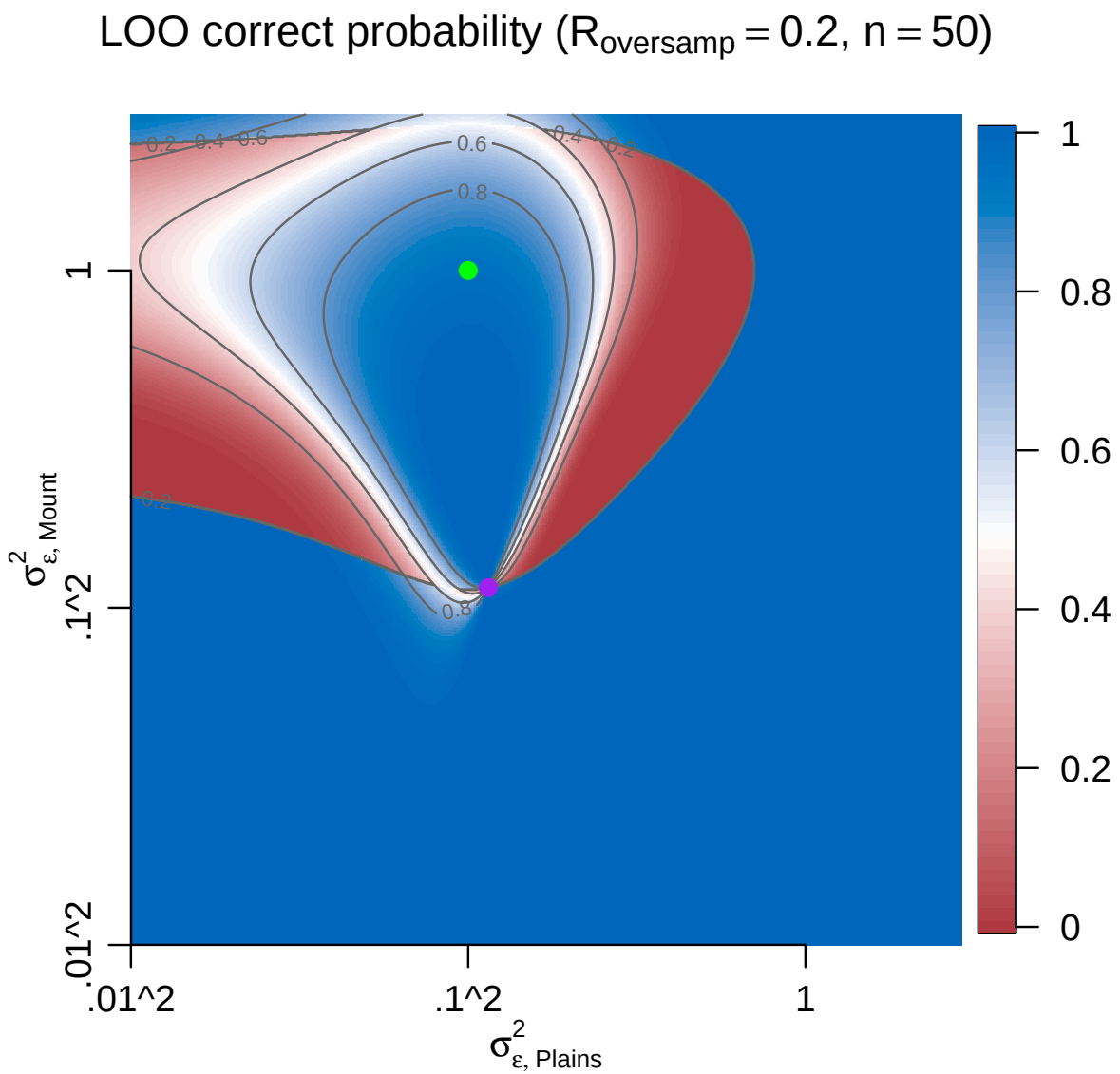
Rabinowicz, A. and S. Rosset (2022). Cross-validation for correlated data. *Journal of the American Statistical Association 117*, 718–731.

Roberts, D. R., V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography 40*, 913–929.

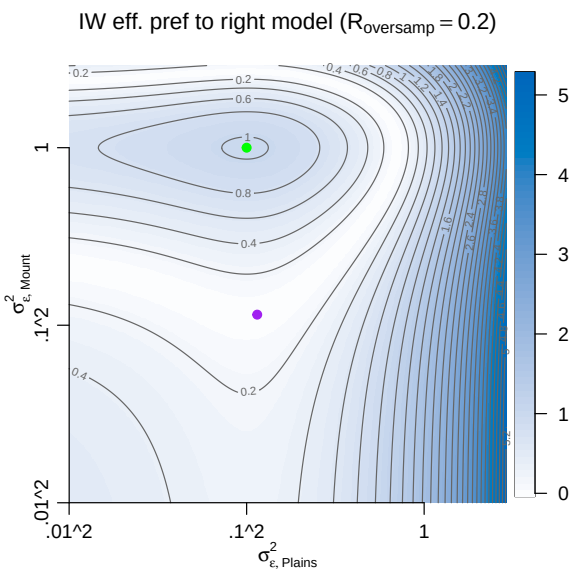
Sugiyama, M., M. Krauledat, and K.-R. Müller (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research 8*.

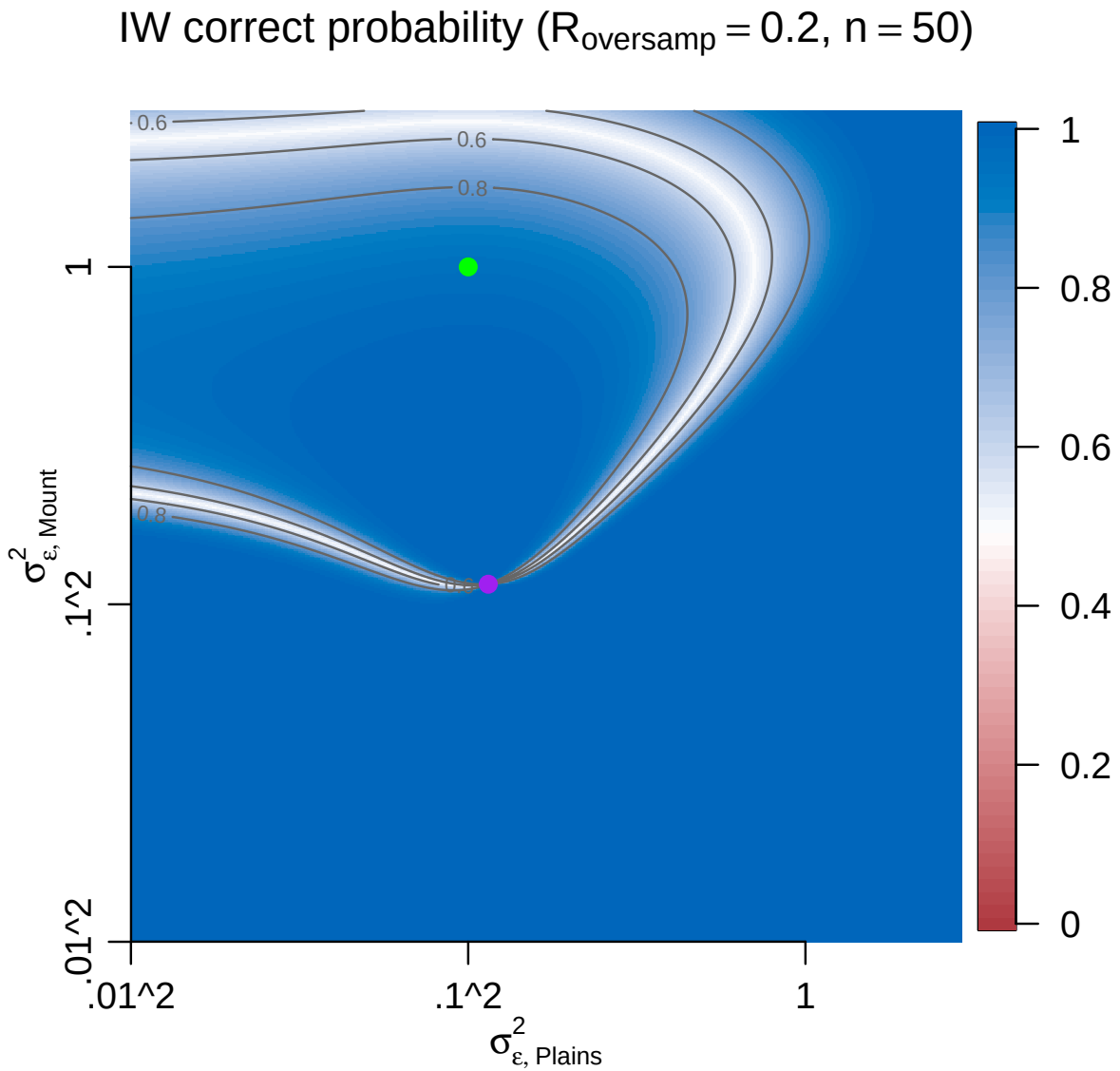
LOO correct model selection probability, $R = 0.2$, $n = 50$



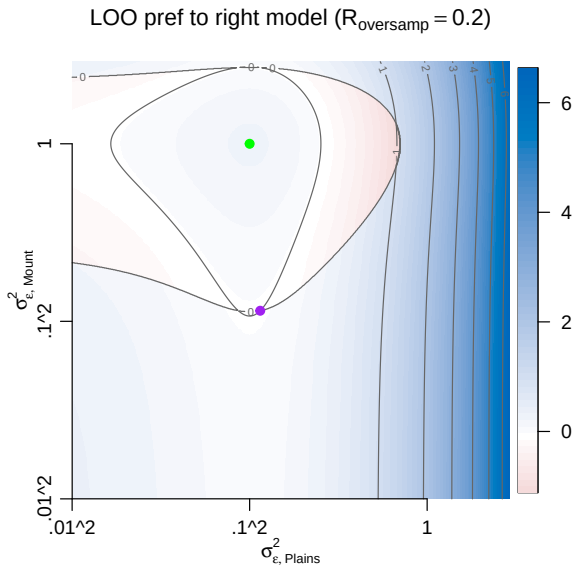


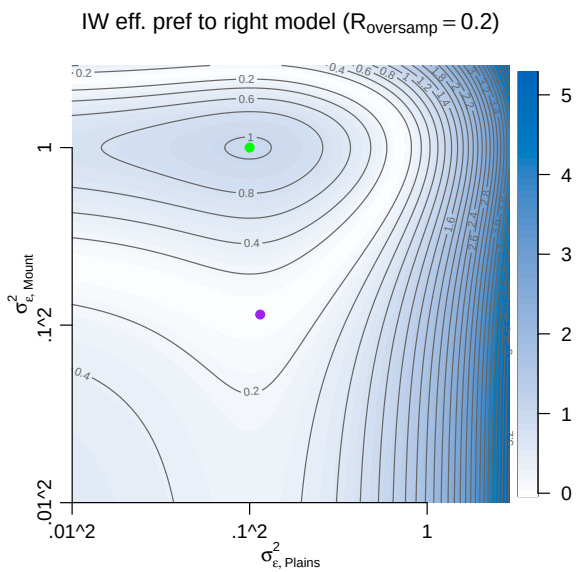
IW correct model selection probability, $R = 0.2$, $n = 50$

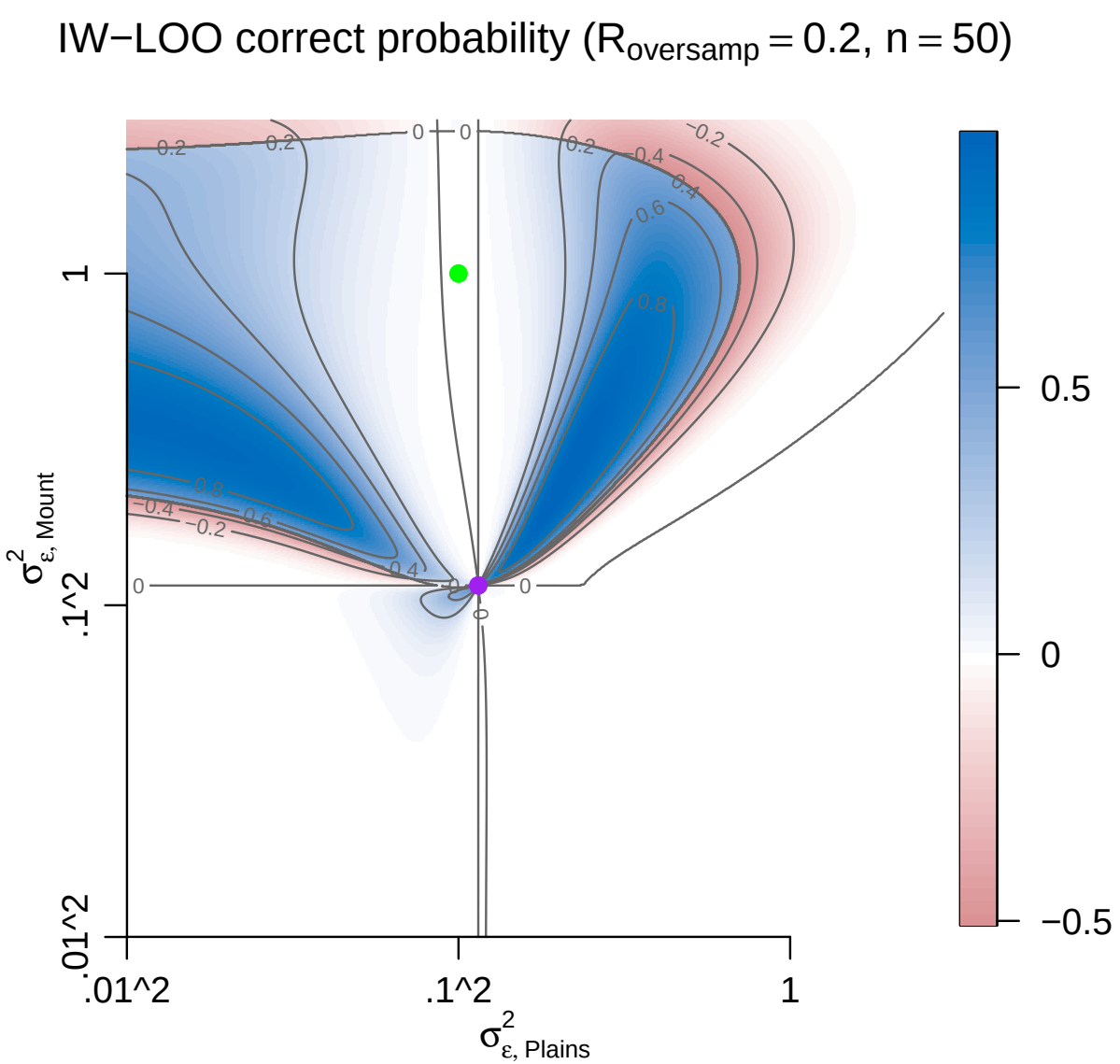




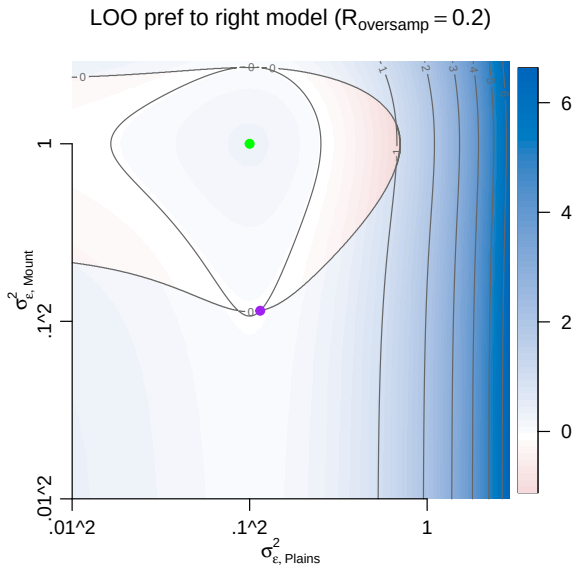
IW-LOO correct selection probability, $R = 0.2$, $n = 50$

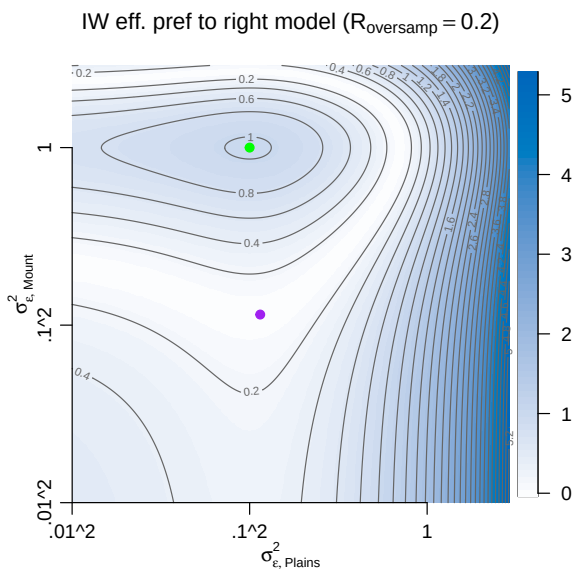


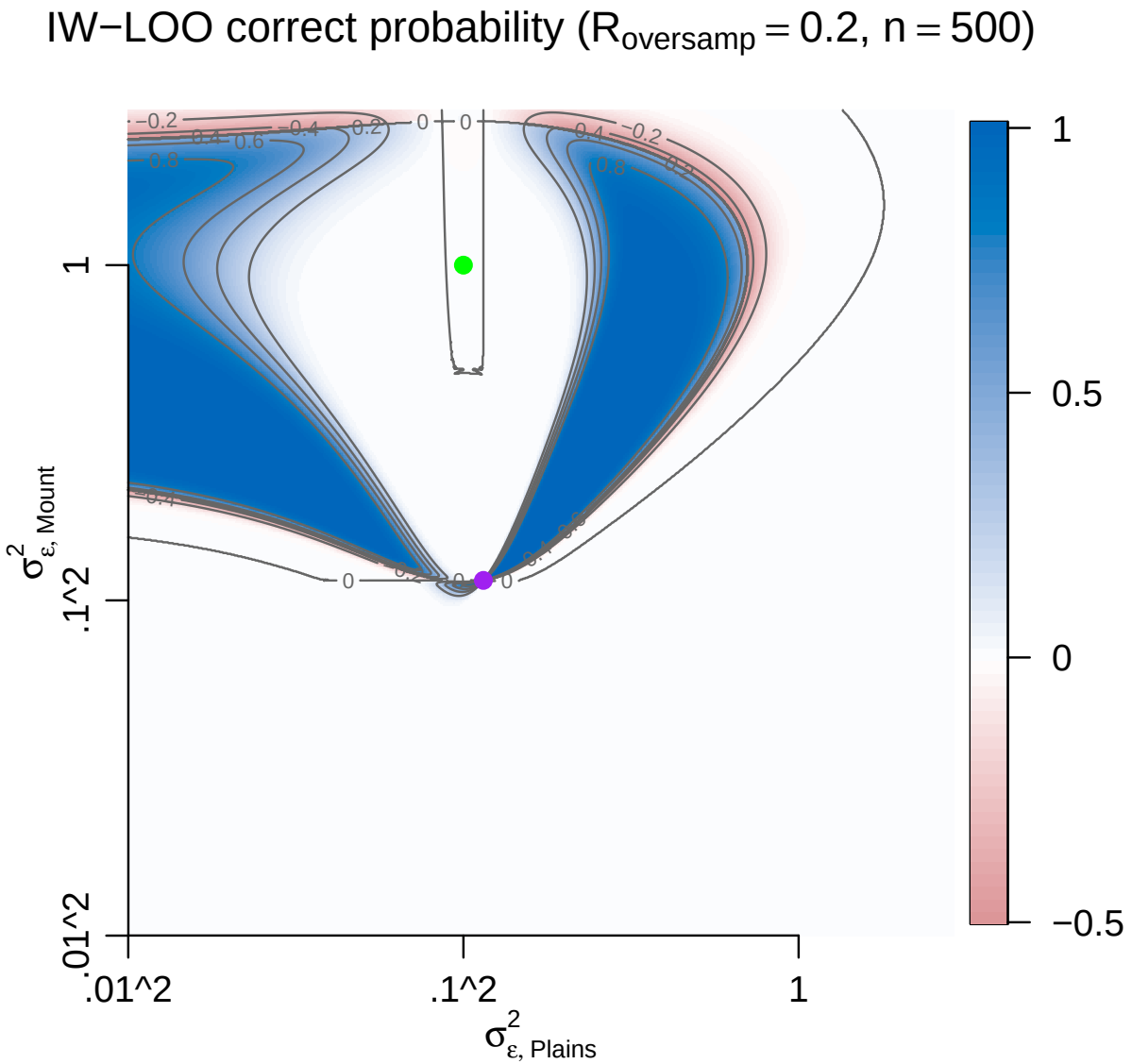




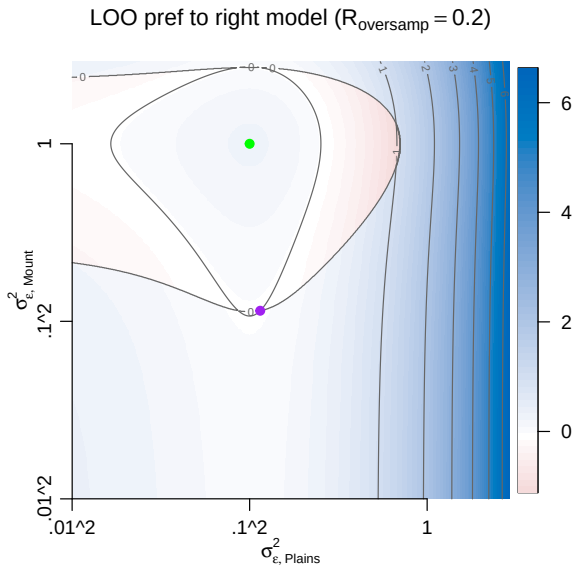
IW-LOO correct selection probability, $R = 0.2$, $n = 500$

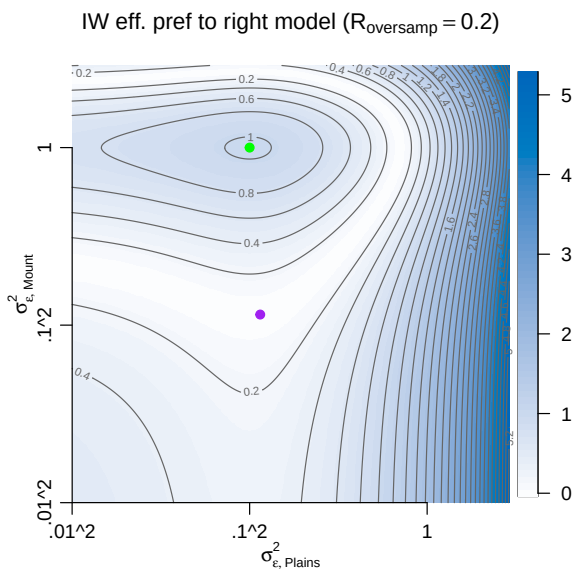


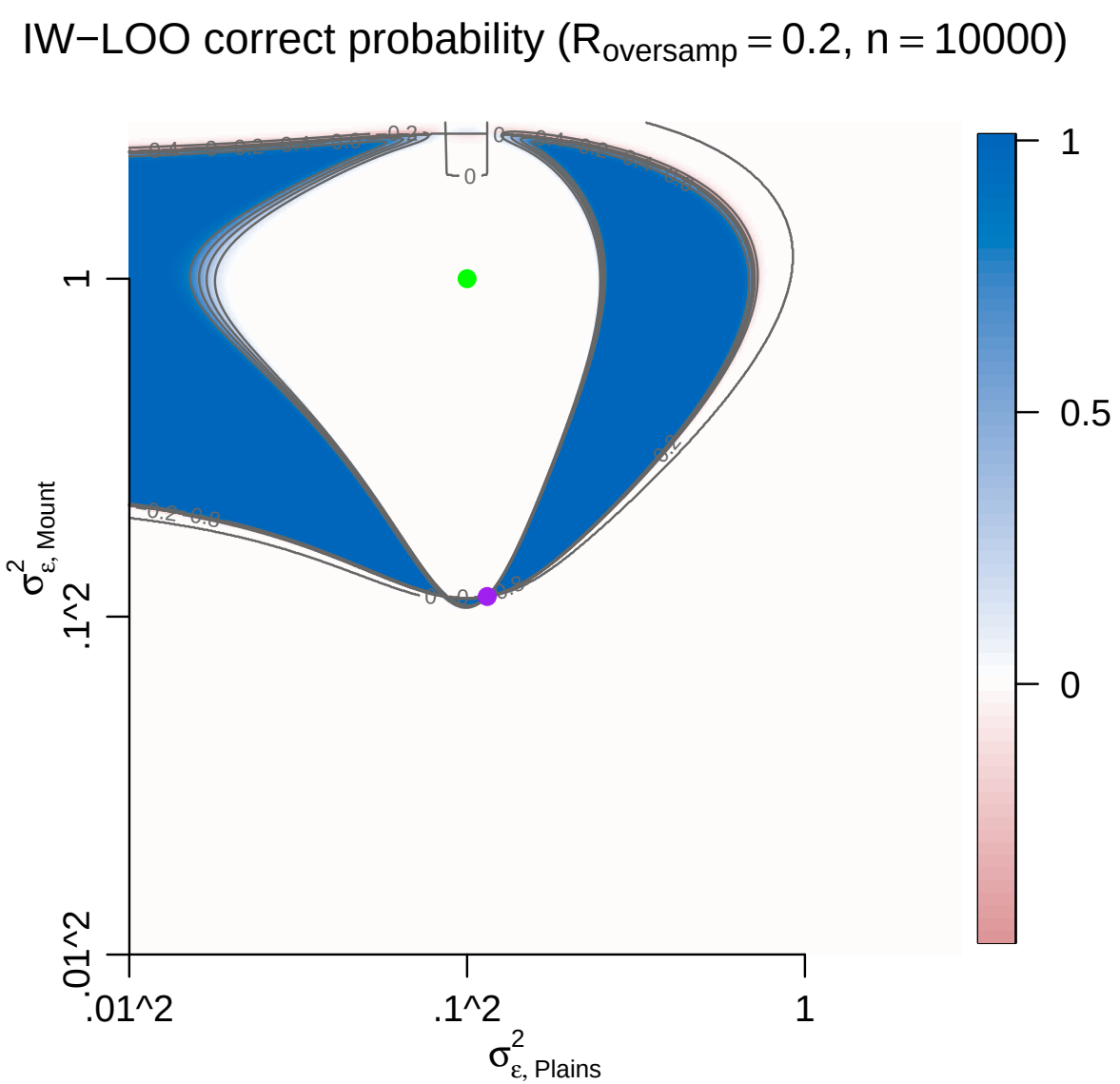




IW-LOO correct selection probability, $R = 0.2$, $n = 10000$







Scoring models

Let

$$\Delta^{\text{LOO}} = \text{LOO}_1 - \text{LOO}_2, \qquad \Delta^{\text{IW}} = \text{IW}_1 - \text{IW}_2, \qquad \text{and}$$

$$\Delta^{\mathcal{D}} = E_{g,g}[\Delta^{\text{IW}}] = E_{f,f}[\Delta^{\text{LOO}}].$$

Under infill asymptotics, use the scores/metrics:

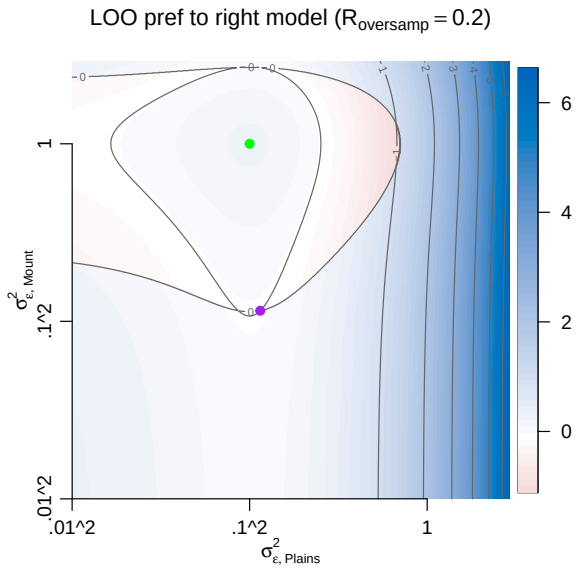
- Probability of selecting the best model
- Mean preference towards best model:

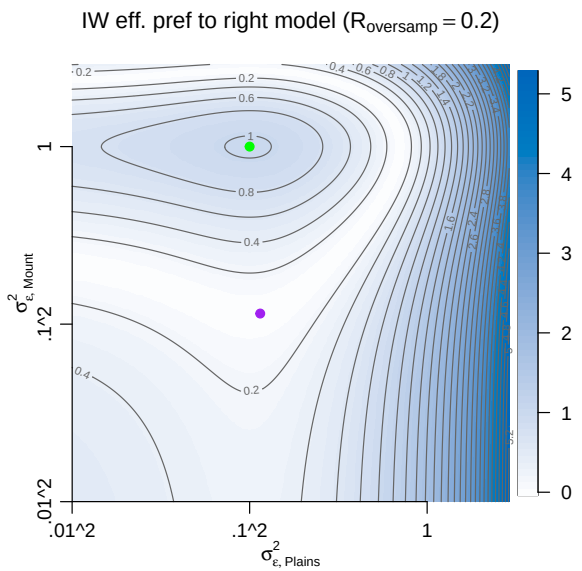
$$|E[\Delta^{\text{LOO}}]|\text{sign}\{E[\Delta^{\text{LOO}}] \cdot E[\Delta^{\mathcal{D}}]\}$$

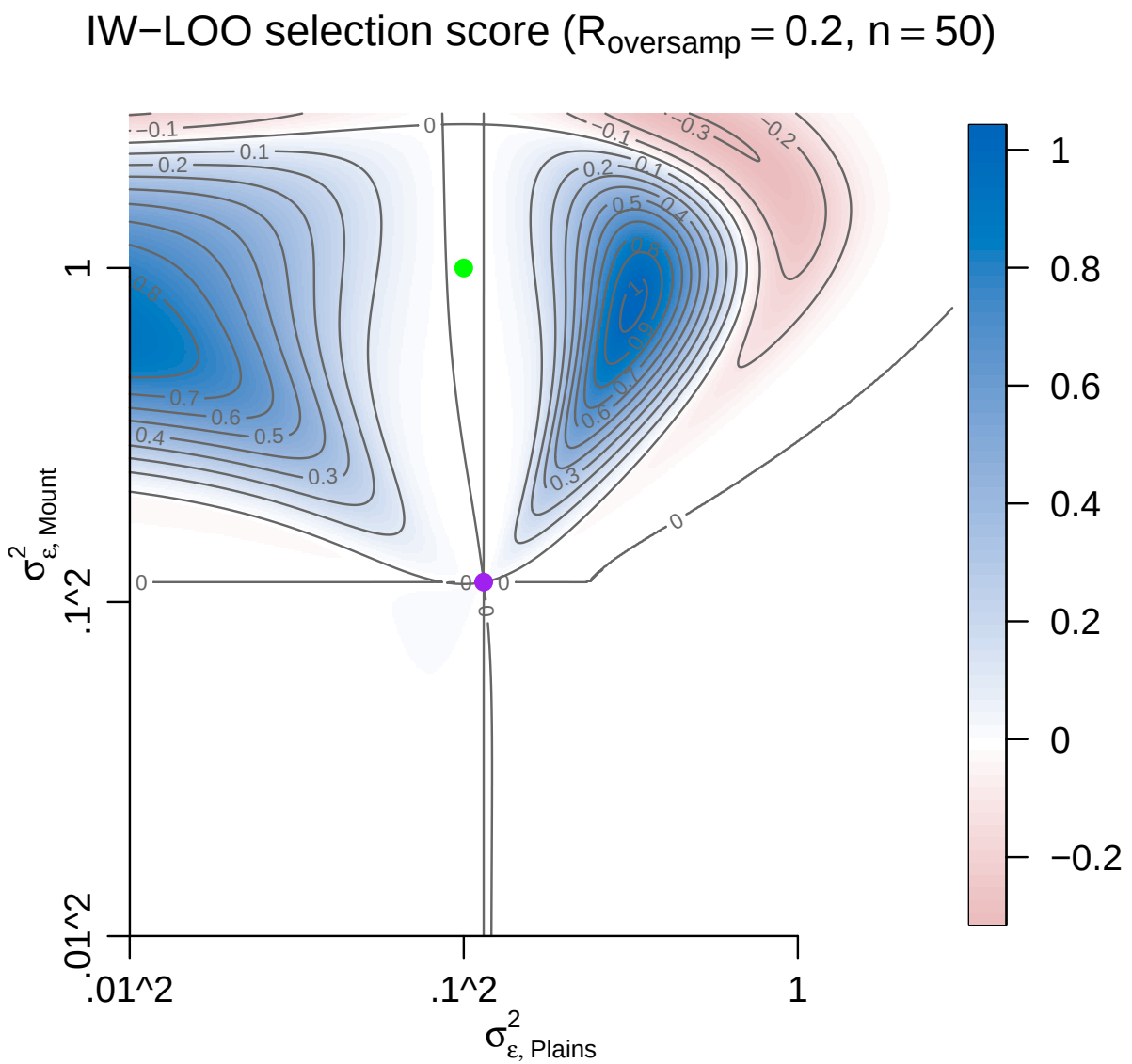
- Expected selection score:

$$E[|\Delta^{\mathcal{D}}|(I_{\text{correct}} - I_{\text{wrong}})]$$

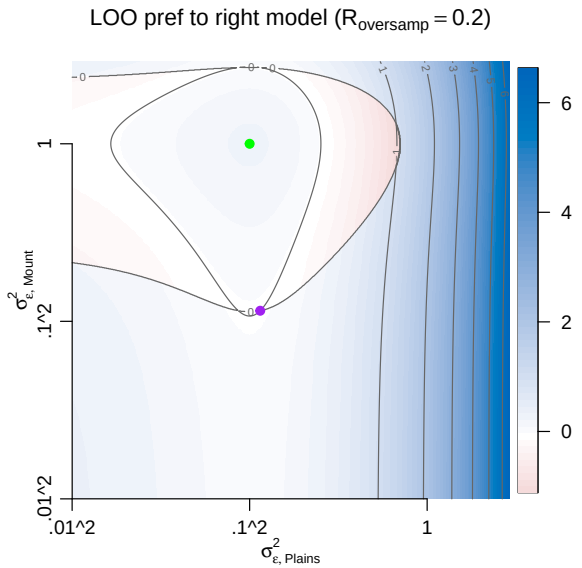
IW-LOO expected selection score, $R = 0.2$, $n = 50$

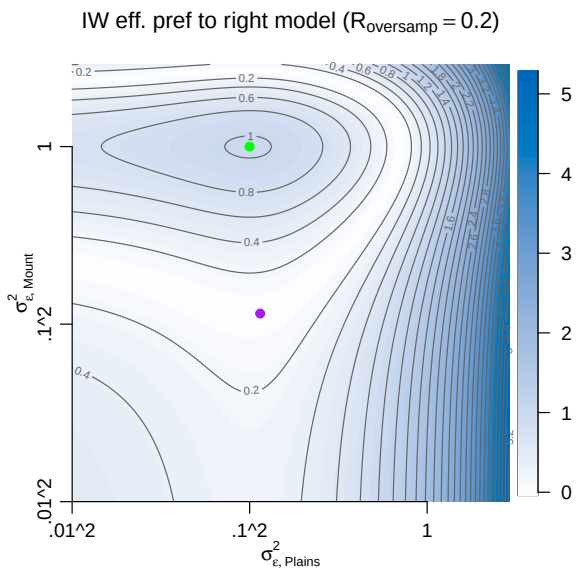


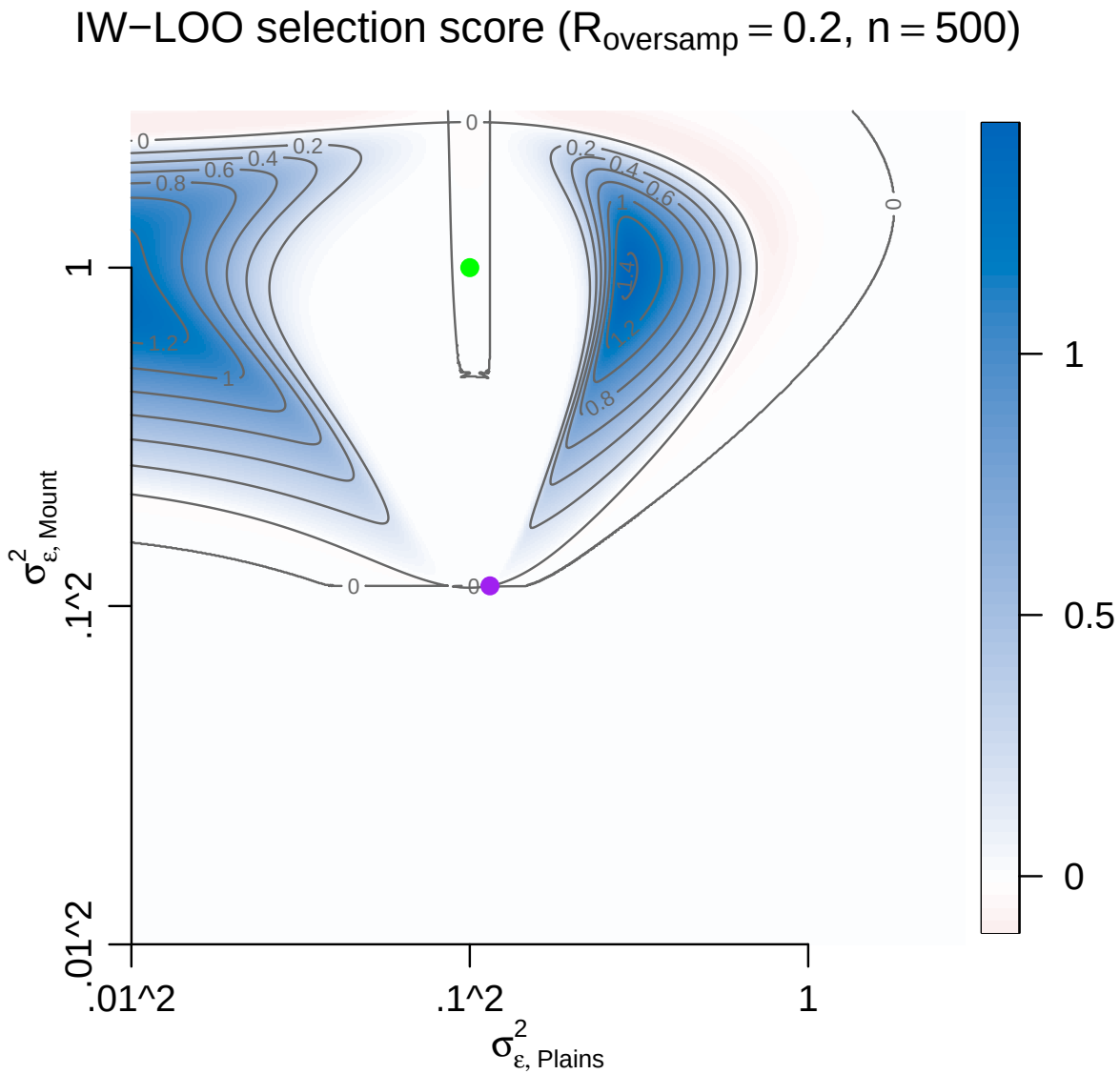




IW-LOO expected selection score, $R = 0.2$, $n = 500$







IW-LOO expected selection score, $R = 0.2$, $n = 10000$